Abstract

SHOCKLEY, KEITH ROPP. Functional genomics investigation of microbial physiology in the hyperthermophilic microorganisms *Pyrococcus furiosus* and T*hermotoga maritima***.** (Under the direction of Dr. Robert M. Kelly).

The recent availability of complete genome sequences for hyperthermophiles has provided an enormous amount of DNA sequence data which, in principle, contains the raw information necessary for the discovery of thermostable biocatalysts of industrial importance. However, the sequence data must be deciphered in order to understand how the cell utilizes its genetic inventory in performing the metabolic, biochemical, and physiological functions necessary to carry out basic cell processes. Since there are currently no genetic systems available to study these organisms, DNA microarrays present an important opportunity to annotate hyperthermophilic genomes through strategic differential expression experiments.

Here, reproducibility of full genome cDNA microarray data was investigated in the hyperthermophilic bacterium *Thermotoga maritima* through the implementation of a highly replicated loop design in conjunction with continuous culture growth. Three mechanical steady states corresponding to two separate dilution rates (D=0.25 hr^{-1} and D=0.17 hr^{-1}) at two distinct temperatures (80^{o}C and 85^{o}C) were assesed. ANOVA-based mixed models were shown to more adequately model transcriptional differences and demonstrated that many genes exhibited a low, but statistically significant, variation in expression within a steady state. Of the 422 ORFs that were differentially regulated between steady states, 93 ORFs also showed significant variability in gene expression within a steady state.

Subsequent to targeted studies of the heat shock response in *P. furiosus* after one hour, full genome analyses of heat shock in *P. furiosus* and *T. maritima* were investigated through time-course experiments. The heat shock response in both of these organisms included induced expression of genes encoding molecular chaperones and proteins involved in the stabilization and repair of DNA. Genes encoding the thermosome (PF1974), the small heat shock protein (PF1882) and a VAT protein (PF1883) were up-regulated in *P. furiosus*, regardless of whether growth was performed on peptides or maltose, while genes encoding two major heat shock operons, *dnaJ-grpE-hrcA* (TM0849-TM0850-TM0851) and *groEL-groES* (TM0505-TM0506), were consistently up-regulated in *T. maritima*. Finally, expression of genes encoding important components of energy metabolism was down-regulated upon thermal stress.

In addition to regulating gene expression due to thermal stress, *P. furiosus* modulated transcription due to differences in media formulation. Genes encoding multiple ABC-type transporters and glycosyl hydrolases were selectively induced when *P. furiosus* was grown on different carbohydrates. In addition, the presence of yeast extract or S^0 in the medium led to differential gene expression of annotated ORFs. Growth on chitin induced the expression of genes encoding an extracellular chitinase (PF1234), an ABC transporter system (PF0357-PF0361), an intracellular chitinase (PF1233), a putative deacetylase (PF0354), a putative glucosaminidase (PF0356), a glucosamine kinase (PF0356) and an isomerase (PF0362).

Transcriptional results presented here indicate that *P. furiosus* and *T. maritima* are able to adapt to a variety of growth environments. These findings provide a foundation for genome-scale analysis of hyperthermophilic organisms and offer a basis for future studies of important features of hyperthermophilic metabolism.

Functional genomics investigation of microbial physiology in the hyperthermophilic microorganisms *Pyrococcus furiosus* and T*hermotoga maritima*

By

Keith R. Shockley

Submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the Degree of Doctor of Philosophy.

Chemical Engineering

Raleigh, NC
May, 2004

APPROVED BY

_____	_____
Dr. David F. Ollis	Dr. Todd R. Klaenhammer
_____	_____
Dr. Ross W. Whetten	Dr. James W. Brown

_____

Dr. Robert M. Kelly
Chair of Advisory Committee

Biography

Keith R. Shockley was born on January 18, 1975 in Columbus, GA. He moved to St. Johnsbury, VT at a young age. At the age of 9, he moved to Las Vegas, NM. While in Las Vegas, he attended Robertson High School, where he pursued interests in baseball and chess; he graduated as class valedictorian, lettered in baseball and received 5th place in the New Mexico State High School Chess Championship his senior year of high school. After high school, he attended New Mexico State University in Las Cruces, NM, where he received a Bachelor of Science degree in Chemical Engineering with a minor in Biochemistry in May of 1998. In the Fall of 1998 he began studies in Chemical Engineering program at North Carolina State University in Raleigh, NC, where he has pursued a Ph.D. in Chemical Engineering with a minor in Biotechnology. After completion of his degree, he plans to enter a postdoctoral training program at The Jackson Laboratory in Bar Harbor, ME, where he will continue to work on the development and implementation of DNA microarrays.

Acknowledgments

This work was supported in part by a GAANN fellowship and grants from the Biotechnology Program, the National Science Foundation and the Energy Biosciences Program, U.S. Department of Energy. I would like to acknowledge the members of my Ph.D. committee, who provided insight, constructive criticism and advice that benefitted this work: Dr. David F. Ollis, Dr. James W. Brown, Dr. Todd R. Klaenhammer and Dr. Ross W. Whetten. I would especially like to thank the chair of my committee, Dr. Robert M. Kelly, for providing invaluable guidance during my time at NCSU. He not only invested the resources and inspiration that made this research possible, but has become a personal role model in my life. My gratitude also extends to present and past members of Dr. Kelly's laboratory, especially Shannon Connors, Clemente Montero, Matthew Johnson, Kevin Epting, Chung-Jung Chou, Donald Ward, Marybeth Pysz, Swapnil Chhabra, Amitabh Sehgal, Lara Chang and Jun Gao for their technical support and advice. Kevin Scott and Dr. Russ Wolfinger (SAS, Cary, NC) contributed greatly to the statistical analyses in this work. Additionally, Dr. Len van Zyl (Array Xpress, Raleigh, NC) provided strategic assistance during my time at NCSU. My family, especially my father, grandma and sister, have always encouraged my academic pursuits and stood by me through all of life's circumstances. I could not have completed this effort without their support and I am grateful for their love and patience. Finally, I would like to acknowledge God for providing the Grace and "moments of light" that brought this project to completion.

Table of Contents

List of Tables

List of Figures

Chapter 1: Applications of cDNA Microarrays to Prokaryotic Systems

Keith R. Shockley and Robert M. Kelly

Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

Microbial genomics

The modern discipline of "microbial genomics" stems from the discovery of principles of heredity proposed by Gregor Mendel in the mid-nineteen century (40) and independently rediscovered through the work of three European scientists at the turn of the twentieth century (14, 16, 69). The next major breakthroughs occurred when "the transforming factor" (or DNA) was determined to be the material responsible for heredity (5) and the structure of DNA was solved (74). Nearly 100 years after Mendel's original discovery, the genetic code was elucidated (45) and recombinant DNA technology was developed (12, 29). The development of automated methods for sequencing DNA eventually set the stage for complete genome sequencing efforts. Sequencing technologies now allow the most productive high-throughput facilities to sequence an entire microbial genome with 8-10 fold coverage in a single day (18).

Beginning in the late 1990's, the availability of complete genome sequences for a range of model microorganisms opened the door for the investigation of living systems at a new level of molecular detail. The first complete genetic map of a free-living organism was performed on the Gram-negative bacterium *Haemophilus influenzae Rd* strain *KW20* whose only natural host is human (23). Currently, genome sequences are accessible for microorganisms that grow at extremely low pH, temperatures above the boiling point of water, pressures greater than 200 atmospheres, elevated metal concentrations, high radiation fluxes and elevated salinity. These complete genome sequence data sets comprise the genetic inventory and regulatory sequences necessary for the array of responses that a given organism has at its disposal to deal with various physiological perturbations and provide locations of open reading frames (ORFs) in a chromosome. In the public domain, genome

sequence information for at least 16 Archaea and 438 Bacteria (9) is now available and many other complete or nearly complete genomes have been fully or partially sequenced in the private sector (see www.tigr.org/tdb/mdb/mdbcomplete.html for current genome sequence number estimates).

The availability of such an enormous amount of microbial genome sequence data led to a shift in emphasis from generating sequence information to studying it. Genomic data sets provide the underlying genetic blueprint of an organism, but comparative genomic analyses may lead to a deeper understanding of phylogenetic relationships among organisms. Comparing the genomes of closely related species provides an opportunity to identify regulatory elements, locate conserved gene clusters, and improve predictions of co-regulated genes (operons and regulons). For instance, regulatory elements may stand out from the 'noise' of sequence data when comparisons are made between close relatives. The genomes of two model hyperthermophilic bacteria, *Aquifex* and *Thermotoga*, have recently been sequenced. These sequences reveal that lateral gene transfer events may have been common between these high-temperature organisms and the Archaea (17, 42, 49). Although widespread, acquisition of bacterial-like genes into archaeal genomes may be more associated with metabolism and transport than more 'informational' processes (21). Comparing genome sequence information may permit revised estimates of the total number of genes present in a given organism and insights into genome evolution. Although quite similar, genome sequence information amongst the Pyrococci indicates a high number of differential gains and losses of genes since they diverged; in fact, *Pyrococcus furiosus* genome is approximately ten percent larger than the genomes of *Pyrococcus abyssi* and *Pyrococcus horikoshii* (36).

Prokaryotic functional genomics

While insights gained directly from genomic data are considerable, they have at the same time highlighted our lack of understanding of how cells function. Advances in molecular biology have allowed a transition into 'modular biology,' in which biological processes of interest (or modules) are studied as multifaceted systems of interacting macromolecules. It has long been known that cells are much more than a dilute bag of diffusing proteins but, rather, are composed of an internal organization containing many complex interactions following time-dependent associations and based upon elaborate molecular scaffolding events. Complete genome sequence information does provide a list of ORFs that an organism has at its disposal to accomplish the necessary cellular activities, but it remains to identify and characterize the relationships and regulation of the genetic elements present in those DNA sequences. To fully appreciate the genetic potential inherent in microbial communities, it will be necessary to understand how genes cooperate to regulate cellular activities. Sophisticated experimental techniques are needed to ascertain how DNA sequence confers function and how epigenetic changes relate to this process. Regulatory elements need to be both identified and understood.

To date, most of the functional genomic research aimed at the prokaryotic community has focused on a select number of pathogenic bacteria, such as *Escherichia coli* and *Clostridia*. It is important to try to understand the connections between the molecular and the physiological in order to identify targets of interest. However, these targets are not isolated entities operating inside of a cell; instead they are components of a biological context involving a highly structured network of cell responses composed of nonlinear connections and feedbacks. Gene expression analysis is an important means to interrogate these

biological conundrums, because changes in the physiology of a cell will necessarily be accompanied by changes in gene expression.

The central dogma of genetics refers to the process of information flow in which DNA is first "transcribed" into RNA and then "translated" into protein (33). The transcriptional products that code for proteins are termed messenger RNAs, or mRNAs. Unique mRNAs lead to the formation of different proteins. Genetic regulation may be described as the process in which a cell decides whether a gene is active or inactive (33, 34). Active genes are those that are being transcribed, whose transcripts are being translated, and whose protein products are actively performing their function. High throughput gene expression can be studied by observing mRNA or protein array maps. However, the majority of gene regulation in bacterial species is most often controlled at the levels of transcription initiation and mRNA turnover. Cells avoid the synthesis of useless intermediates such as unused mRNA or incomplete proteins (41, 68). Therefore, an understanding of the mechanisms that regulate gene transcription is essential to the knowledge of underlying biological processes. Because they lack a true nuclear membrane, transcription and translation can occur at the same time in prokaryotes. Both Bacteria and Archaea contain polycistronic regions (gene clusters) that are often regulated by a single promoter (6, 37).

Approaches to assess *in vitro* transcript differences between two or more populations are based on either differential screening or differential hybridization strategies. Differential screening methods compare the abundance of randomly identified, single transcript, clones between two different cDNA libraries (39). In contrast, differential hybridization schemes utilize the fact that complimentary polynucleotide sequences can bind to each other much more strongly than to dissimilar polynucleotides (39). Numerous advances allow the

identification and quantification of differentially expressed genes, including Northern and dot blot analyses (3, 22), nuclease protection assays (7), cDNA library screening (35, 60), differential display of RNA (39, 78), quantitative PCR (47), serial analysis of gene expression (71) and DNA microarrays (55).

DNA microarrays – Quantitative tools to study prokaryotic systems

The most powerful tool for studying transcription patterns is the DNA microarray. The novelty of the DNA microarray assay consists primarily in its parallelism, miniaturization, multiplexing, and automation. The DNA microarray allows the analysis of gene expression in parallel with the direct readout of hybridization results, bypasses the complications associated with additional amplification and sequencing steps, and utilizes a rigid surface substrate that is impermeable to liquids (19, 55, 62, 76). Until the advent of this technology, it was only possible to characterize gene expression a few genes at a time. Now, however, DNA microarrays allow the simultaneous and quantitative analysis of the expression of thousands of known and unknown genes.

Two basic types of DNA microarrays are used most often to study prokaryotic systems: the oligonucleotides microarray and the cDNA microarray. Oligonucleotide arrays may be constructed through *in situ* synthesis, or the deposition of previously synthesized oligonucleotides (usually less than 60 bp) onto silicon wafers (62, 76). One end of each oligonucleotide is covalently attached to the solid substrate while the other is freely accessible to hybridize to targets present in solution, allowing maximal hybridization (62). In order to construct cDNA microarrays, nucleic acid probes (usually 0.5kb – 1Kb) are usually PCR amplified directly from genomic DNA. Each element of an array is generated

by robotically depositing nanoliter volumes of probe DNA solutions onto either poly-L-lysine or amino-silane treated glass microscope slides, followed by chemical or heat-treated attachment to the surface and denaturation (10, 19, 55). Probes may be bound either through covalent or ionic attachment, depending on the nature of the slide coating. Robotic spotting is usually performed from 96- or 384-well plates through the use of the standard "quill" pins, solid pins, or the Pin-Ring device produced by Genetic Microsystems, which is now Affymetrix in Santa Clara, CA (10, 76). Robots are capable of printing in 0.005 μl increments and 500 μm spacing (55). Most pins draw liquid through capillary action and spot the probes onto the slide through the combined effects of the downward motion of the pin and the surface tension associated with the slide (10).

The initial step in prokaryotic DNA microarray procedures is the isolation and purification of total RNA. In order for the system to function properly, the sample RNA pools must be very pure because cellular protein, lipid, and carbohydrate can mediate nonspecific binding to slide surfaces (20, 76). Extracted RNA is then reverse-transcribed into complementary DNA, or cDNA. Labeling may occur during transcription through direct incorporation of a fluorescent dinucleotide base into the cDNA sequence, or after transcription by first generating cDNA with a modified dinucleotide base and subsequent chemical reaction with an appropriate fluorophor. The most common labeling procedures utilize the fluorescent molecules Cy3 and Cy5. The DNA microarray relies on the concept that, under sufficiently stringent conditions, labeled cDNA molecules (or targets) will selectively hybridize to its complimentary DNA sequences (or probes) that are prefabricated onto a solid substrate support. The probes for cDNA microarrays may consist of PCR products, randomly chosen library cDNAs, or short oligonucleotide sequences. Scanning the

slide regions that possess the desired probe sequences at wavelengths corresponding to the absorption spectrum of the fluorophor produces signal intensities, which serve as measures of the relative abundance of each target sequence in an experimental sample. In this way, the level of expression can be estimated for each probe attached onto the array. The signal intensity data obtained from DNA microarrays provides both static information relating which genes are expressed under a given set of tested conditions and information about how the pattern of gene expression relates to the patterns found in other genes (19, 55).

The first cDNA microarray reported in the literature (55) was a targeted DNA microarray (i.e., a microarray containing a subset of the full genome) in which total signal intensities were used to normalize between channels and simple fold changes were used to describe differential expression of selected targets. While this straightforward method for handling the data is able to reveal large changes in transcript abundance between two samples, current studies focus on data sets that result from whole genomes and often employ more complex procedures for data normalization and analysis that allow a higher resolution picture of transcriptional changes over multiple treatments. For instance, the normalization step of ANOVA-based mixed model approaches remove error due to global effects associated with treatments and dyes (fixed effects) while gene specific models remove contributions arising from slide regions, pins or spots (random effects). After data normalization, a second ANOVA-based mixed model can then be used to normalize the data set for gene specific effects, which allow individual genes to have different expression level variances. The least square means (lsm) estimates that result from the gene-specific model can be presented in forms analogous to fold change or clustered directly (or in a standardized form) in order to visualize the expression behavior of all the members of a regulon.

Prokaryotic biotechnology

Whether it is playing a fundamental role in maintaining the delicate ecological balance of the earth or maintaining the necessary symbiotic interactions, microorganisms have a major impact on human lives. Not only is prokaryotic life responsible for sustaining biogeochemical cycles, but they also are essential to the healthy development of plants and animals. Microbes can process virtually every one of the 92 naturally occurring elements in the periodic table, including plutonium, which make them highly sought tools of bioremediation efforts. Sources of bio-relevant chemicals have been extended to the coldest artic tundras, the harshest deserts, and in the bottoms of the deepest oceans (13). Nevertheless, even though microorganisms thrive in almost any imaginable niche on this planet, less than 1% of known microbes can be cultured in the laboratory and accepted estimates predict that more than 98% of microbial life has yet to be discovered. Therefore, in spite of the many advances that have taken place within microbial technology, there still exists a wealth of microbial genetic information on Earth that has the potential to revolutionize the biotechnology industry. Given this potential, and the successes in the history of microbial biotechnology, it seems beyond argument that microorganisms still have much to offer modern technological research endeavors. Now it is a matter of accessing this potential.

Genomic information from microbes has already impacted biological, biomedical, bioengineering and biomaterials research. Each of four major market segments of biotechnology (biomedical, agricultural, environmental and industrial sectors) depends on technology derived from microbial systems. Through advances in metabolic engineering and directed molecular evolution, efficient and selective microbial biocatalysts are currently used

in the manufacture of items such as food and agricultural products, detergents, textiles, commodity chemicals, pharmaceutical intermediates and drug substances. It was recently noted that 61% of drugs and drug leads (1981-2002) can be traced to or derived from natural products (43), including 78% of antibacterials and 74% of anticancer agents. Many of these products must be generated by the microbial "cell factory" systems, because these products are so complex that they are too expensive to produce or cannot be synthesized in the laboratory. Industrially relevant bioactive chemicals may be derived from the natural forms or as chemical derivatives that are needed during primary metabolism (essential for cell growth) or the secondary metabolism that perform other functions of these organisms, such as those that are of interest in bioremediation efforts.

In past years, novel enzymes or chemicals were discovered through directly purifying the protein of interest from cell extracts. However, many of the advances in biotechnology that have occurred in recent years have been the result of the large expansion in the number of DNA sequences available. With the aid of this information, cloning and expression trials resulted in a few enzymatic activities of interest. However, full genome sequence information has allowed an even greater expansion of data and opened up an unprecedented assemblage of information to microbial biotechnology.

Hyperthermophiles

The diversity of extremophilic microorganisms on this planet should be considered in terms of the genes that making up these species as well as the total numbers of sub-species comprising this group. Like other known groups of microbes, hyperthermophiles can sense and respond to their environments in rich ecological niches that involve symbiotic

relationships (27), virus mediated genetic events (50), and selective transport mechanisms (11). The availability of complete genome sequences for hyperthermophilic microorganisms promises to revolutionize the study of these biological systems and expedite the identification of enzymes and biotransformations of biotechnological significance. Because of their relatively small genome sizes (~2 Mb) and the substantial amount of data available from previous work in characterizing their biochemical and structural features, hyperthermophiles also present an important opportunity to comprehensively annotate genomes through strategic differential expression studies.

Hyperthermophiles (optimal growth temperatures of 80°C and higher) have been isolated from geographically-diverse hydrothermal environments that range from deep sea vents to terrestrial hot springs (28, 52, 64, 66). Almost all of them are strict anaerobes, and the majority are obligate heterotrophs that require the reduction of elemental sulfur (S^0) to H_2S (1, 31). Of the more than twenty genera of hyperthermophiles currently known, only *Aquifex* and *Thermotoga* are members of the Bacteria (1). The remaining genera belong to the Archaea, and life above 95^oC is confined solely to the Archaea (15, 64).

The demand for thermostable enzymes arises from the fact that most industrial enzyme processes still utilize mesophilic enzymes for catalysis even though the reactions are generally operated at elevated temperatures (65, 77). Most of the native thermostable enzymes purified for these industrial applications are derived from *T. maritima*, *P. furiosus*, or closely related strains, and are active at the extreme conditions found in industrial operations (2, 28, 44). These enzymes, including dehydrogenases, DNA polymerases, amylases, and proteases, are currently used in the food, chemical, and biochemical research industries (28, 77). The DNA polymerase from *Thermus aquaticus* brought about the

polymerase-chain-reaction (PCR) technology (2, 28). DNA polymerases isolated from *P. furiosus* provide "proofreading" ability and are now commercially available (28). The prospect for new enzymes is great considering that there are known enzymatic conversions for which no genes are currently known and a variety of expected metabolic pathways appear to be missing altogether in these high temperature organisms.

Prokaryotic studies

DNA microarrays have been used for such experimental goals as gene expression profiling, clinical diagnostics, environmental monitoring and drug discovery. Genome sequences encode numerous known and putative genes that enable the organisms to respond to the environment and much of the microarray research has centered on response to environmental perturbations. As shown in Table 1, DNA microarray analyses (or "transcriptomics") have been used in order to study wide-scale gene expression patterns in many different microorganisms, including *Bacillus subtilus* (e.g., ref. (4, 25, 63)), *Escherichia coli* (e.g., ref. (46, 53)), *Pseudomonas aeruginosa* (e.g., ref. (56, 72, 75), *Thermotoga maritima* (11), and *Pyrococcus furiosus* (57-59). However, most large-scale transcriptional analyses have been limited to the mesophilic bacteria. Microarrays have been used to study such bacterial processes as osmotic stress (30), heat shock (54), acid stress (70), oxidative stress (79), DNA damage (32), drug resistance (61), nutrient limitation (48), starvation (8), carbon utilization (11) and biofilm development (75). These studies have yielded many candidates for differential gene expression and improved annotation that have remarkably improved our understanding of cell physiology and should enhance efforts devoted to microbial biotechnology. It is evident that microbial genomics presents

researchers with an enormous resource of genetic potential; at present, about 50% of each new genome sequence is comprised of genes that have unknown function (24) and approximately 40% of the predicted ORFs in new genomes are unique. Although many different biological phenomena have been investigated through transcriptional profiling, growth temperature and medium composition have been the focus of much of the DNA microarray-based investigations.

Growth temperature variation

DNA microarrays have been used to study the transcriptional response of microorganisms to variations in temperature. Since elevated temperatures provoke much transcriptional change, most of these studies have focused on effects of increased temperature. The first microarray-based analysis of transcriptional response due to heat shock focused on the enteric, Gram-negative bacterium *Escherichia coli* (54), that controls the heat shock response through the activity of the alternate sigma factors σ^{32} and σ^{E}. Results from this study identified novel components of the heat shock response and supported the combined work from previous studies extending back many years; the expression of genes encoding 23 previously identified members of the heat shock stimulon was dramatically induced due to a temperature upshift from 37^{o}C to 50^{o}C (54). In contrast to *E. coli*, microarray analysis of heat shock in *B. subtilis* confirmed that many heat shock genes were controlled by the secondary sigma factor σ^{B} (25). Monitoring the induction of heat shock gene expression over the course of 20 minutes led to identify approximately 70 additional members of the heat shock regulon in this organism (25). The effect of increasing temperature, over a 50 minute time span, on transcriptional response in the foodborne

pathogen *Campylobacter jujuni* was studied with DNA microarrays (67). While the expression of core heat shock genes was stimulated by increased temperature, the up-regulation of genes involved in membrane makeup indicated that *C. jujuni* had a different protein membrane composition at different temperatures (67).

Because hyperthermophiles must, by definition, live in very extreme environments, they represent a unique opportunity to discover how organisms respond to thermal stress in their natural environments. Targeted DNA microarray analyses of the archaeon *Pyrococcus furious* (59) and the hyperthermophilic bacterium *Thermotoga maritima* (51), indicated that while molecular chaperones are critically important to stability of protein structures in these organisms, ATP-dependent proteolysis may play a lesser role in these high-temperature microorganisms than has been seen in mesophilic organisms.

Medium composition

Many microarray-based studies of microbial systems have studied differences in nutritional conditions available to the organisms. Microorganisms depend on their ability to adapt to changing availability of nutrients for their survival. DNA microarrays were able to detect the up- and down-regulation of many genes in *Escherichia coli* grown in different media containing glucose, glycerol or acetate as a primary carbon source (46). Transcriptional profiling of *Bacillis subtilis* indicates that the presence of sulfate or methionine as the sole source of sulfur results in the differential expression of genes involved in the S-box regulon and transporters *yhcL*, *ytmJKLMN* and *yxeMO* (4). Whole-genome DNA microarray analysis of *P. furiosus* indicated that the organism possessed extensive coordinate regulation of transcription in response to proteolytic or maltose-based growth (57). In *P.*

furiosus, genes involved in maltose transport and the biosynthesis of 12 amino acids, ornithine and citric acid cycle intermediate products when grown on maltose; growth on peptides led to the up-regulation of genes involved with the production of acyl and aryl acids and 2-keto acids (57). Interestingly, despite being classified as a phylogenetically primitive organism, microarray-based transcriptional analysis of the hyperthermophilic bacterium *Thermotoga maritima* has versatile and discriminating mechanisms for utilizing complex carbohydrates (11). In fact, the transcriptional profiles indicated that *T. maritima* is able to recognize sugar backbone and linkages in the utilization of complex carboyhydrate substrates (11).

References

1. **Adams, M. W. W.** 1999. The biochemical diversity of life near and above 100°C in marine environments. J. Appl. Microbiol. Sym. Suppl. **85:**108S-117S.
2. **Adams, M. W. W. a. K., R. M.** 1995. Enzymes from microorganisms in extreme environments. Chem. Engr. News **73:**32-42.
3. **Alwine, J. C., D. J. Kemp, and G. R. Stark.** 1977. Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes. Proc. Natl. Acad. Sci. USA **74:**5350-5354.
4. **Auger, S., A. Danchin, and I. Martin-Verstraete.** 2002. Global expression profile of *Bacillus subtilis* grown in the presence of sulfate or methionine. J. Bacteriol. **184:**5179-5186.
5. **Avery, O. T., MacLeod, C.M. and McCarty, M.** 1944. Studies of the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* Type III. J. Exp. Med. **79:**137-158.
6. **Baumann, P., S. A. Qureshi, and S. P. Jackson.** 1995. Transcription: new insights from studies on Archaea. Trends Genet. **11:**279-283.
7. **Berk, A. J., and P. A. Sharp.** 1977. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. Cell **12:**721-32.
8. **Betts, J. C., P. T. Lukey, L. C. Robb, R. A. McAdam, and K. Duncan.** 2002. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. Mol. Microbiol. **43:**717-731.

9. **Burja, A. M., S. Dhamwichukorn, and P. C. Wright.** 2003. Cyanobacterial postgenomic research and systems biology. Trends Biotechnol. **21:**504-511.

10. **Cheung, V. G., M. Morley, F. Aguilar, A. Massimi, R. Kucherlapati, and G. Childs.** 1999. Making and reading microarrays. Nat. Genet. **21:**15-19.

11. **Chhabra, S. R., K. R. Shockley, S. B. Conners, K. L. Scott, R. D. Wolfinger, and R. M. Kelly.** 2003. Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima*. J. Biol. Chem. **278:**7540-7552.

12. **Cohen, S. N., Chang, A. C., Boyer, H. W. and Helling, R. B.** 1973. Construction of biologically functional bacterial plasmids *in vitro*. Proc. Natl. Acad. Sci. USA **70:**3240-3244.

13. **Colwell, R. R.** 2002. Fulfilling the promise of biotechnology. Biotechnol. Adv. **20:**215-28.

14. **Correns, C.** 1900. G. Mendels Regel über das Verhalten der Nachkommenschaft der Rassenbastarde. Berichte der Deutschen Botanischen Gesellschaft **18:**158-168.

15. **Daniel, R. M., and D. A. Cowan.** 2000. Biomolecular stability and life at high temperatures. Cell Mol. Life Sci. **57:**250-264.

16. **De Vries, H.** 1900. Sur la loi de disjonction des hybrides. Comptes Rendus de l'Academie des Sciences Paris **130:**845-847.

17. **Deckert, G., P. V. Warren, T. Gaasterland, W. G. Young, A. L. Lenox, D. E. Graham, R. Overbeek, M. A. Snead, M. Keller, M. Aujay, R. Huber, R. A. Feldman, J. M. Short, G. J. Olsen, and R. V. Swanson.** 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature **392:**353-358.

18. **Drell, D.** 2002. The Department of Energy microbial cell project: A 180 degrees paradigm shift for biology. Omics **6:**3-9.

19. **Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent.** 1999. Expression profiling using cDNA microarrays. Nat. Genet. **21:**10-14.

20. **Duggan, D. J., M. Bittner, Y. Chen, P. Meltzer, and J. M. Trent.** 1999. Expression profiling using cDNA microarrays. Nat. Genet. **21:**10-4.

21. **Faguy, D. M., and W. F. Doolittle.** 1999. Lessons from the *Aeropyrum pernix* genome. Curr. Biol. **9:**R883-R886.

22. **Farrell, R. E.** 1998. RNA Methodologies. Harcourt Brace & Company, San Diego.

23. **Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, and et al.** 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science **269:**496-512.

24. **Fraser, C. M.** 2002. Microbial genome sequencing: prospects for development of novel vaccines and anti-microbial compounds. Sci. World J. **2 Suppl 2:**1-2.

25. **Helmann, J. D., M. F. Wu, P. A. Kobel, F. J. Gamo, M. Wilson, M. M. Morshedi, M. Navre, and C. Paddon.** 2001. Global transcriptional response of *Bacillus subtilis* to heat shock. J. Bacteriol. **183:**7318-7128.

26. **Hemsley, C., E. Joyce, D. L. Hava, A. Kawale, and A. Camilli.** 2003. MgrA, an orthologue of Mga, Acts as a transcriptional repressor of the genes within the rlrA pathogenicity islet in *Streptococcus pneumoniae*. J. Bacteriol. **185:**6640-6647.

27. **Huber, H., M. J. Hohn, R. Rachel, T. Fuchs, V. C. Wimmer, and K. O. Stetter.** 2002. A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. Nature **417:**63-7.

28. **Huber, H. a. S., K. O.** 1998. Hyperthermophiles and their possible potential in biotechnology. J. Biotechnol. **64:**39-52.

29. **Jackson, D. A., Symons, R. H., and Berg, P.** 1972. Biochemical method for inserting new genetic information into DNA of Simian Virus 40: circular SV40 DNA molecules containing lambda phage genes and the galactose operon of *Escherichia coli*. Proc. Natl. Acad. Sci. USA **69:**2904-2909.

30. **Kanesaki, Y., I. Susuki, S. I. Allakhverdiev, K. Mikami, and N. Murata.** 2002. Salt stress and hyperosmotic stress regulate the expression of different sets of genes in *Synechocystis sp* PCC 6803. Biochem. Biophys. Res. Commun. **290:**339-348.

31. **Kengen, S. W. M., Stams, A. J. M., and de Vos, W. M.** 1996. Sugar metabolism of hyperthermophiles. FEMS Microbiol. Rev. **18:**119-137.

32. **Khil, P. P., and R. D. Camerini-Otero.** 2002. Over 1000 genes are involved in the DNA damage response of *Escherichia coli*. Mol. Microbiol. **44:**89-105.

33. **Klug, W. S. a. C., M. R.** 2000. Concepts of Genetics, Sixth ed. Prentice-Hall, Inc., New Jersey.

34. **Kornberg, R. D.** 2000. Eukaryotic transcriptional control. Trends Biochem. Sci. **24:**M46-M49.

35. **Lassner, D.** 1995. Synthesis of cDNA, p. 65-70. *In* T. Kohler, Labner, D., Rost, A. -K., Thamm, B., Pustowoit, B., and Remke, H. (ed.), Quantitation of mRNA by Polymerase Chain Reaction: Nonradioactive PCR Methods. Springer-Verlag, Berlin.

36. **Lecompte, O., R. Ripp, V. Puzos-Barbe, S. Duprat, R. Heilig, J. Dietrich, J. C. Thierry, and O. Poch.** 2001. Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Res. **11:**981-993.

37. **Leigh, J. A.** 1999. Transcriptional regulation in Archaea. Curr. Opin. Microbiol. **2:**131-134.

38. **Loos, A., C. Glanemann, L. B. Willis, X. M. O'Brien, P. A. Lessard, R. Gerstmeir, S. Guillouet, and A. J. Sinskey.** 2001. Development and validation of corynebacterium DNA microarrays. Appl. Environ. Microbiol. **67:**2310-2318.

39. **Matz, M. V., and S. A. Lukyanov.** 1998. Different strategies of differential display: areas of application. Nucleic Acids Res. **26:**5537-5543.

40. **Mendel, G.** 1866. Verhandlungen des naturforschenden Vreines. Abhandlungen **4:**3-47.

41. **Neidhardt, F. C., Ingraham, J. L., and Schaechter, M.** 1990. Physiology of the bacterial cell: a molecular approach. Sinauer Associates, Inc., Sunderland, Massachusetts.

42. **Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, C. M. Fraser, and et al.** 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature **399:**323-329.

43. **Newman, D. J., G. M. Cragg, and K. M. Snader.** 2003. Natural products as sources of new drugs over the period 1981-2002. J. Nat. Prod. **66:**1022-1037.

44. **Niehaus, F., Bertoldo, C., Kahler, M., and Antranikian, G.** 1999. Extremophiles as a source of novel enzymes for industrial application. Appl. Environ. Microbiol. **51:**711-729.

45. **Nirenberg, M. W.** 1963. The Genetic Code. Sci. Am. **208:**80-94.

46. **Oh, M. K., and J. C. Liao.** 2000. Gene expression profiling by DNA microarrays and metabolic fluxes in *Escherichia coli*. Biotechnol. Prog. **16:**278-286.

47. **Orlando, C., P. Pinzani, and M. Pazzagli.** 1998. Developments in quantitative PCR. Clin. Chem. Lab Med. **36:**255-269.

48. **Paustin, M. L., B. J. May, and V. Kapur.** 2002. Transcriptional response of *Pasteurella multocida* to nutrient limitation. J. Bacteriol. **184:**3734-3739.

49. **Pennisi, E.** 1999. Is it time to uproot the tree of life? Science **284:**1305-1307.

50. **Prangishvili, D., S. V. Albers, I. Holz, H. P. Arnold, K. Stedman, T. Klein, H. Singh, J. Hiort, A. Schweier, J. K. Kristjansson, and W. Zillig.** 1998. Conjugation in archaea: frequent occurrence of conjugative plasmids in *Sulfolobus*. Plasmid **40:**190-202.

51. **Pysz, M. A., D. E. Ward, K. R. Shockley, C. I. Montero, S. B. Conners, M. R. Johnson, and R. M. Kelly.** 2004. Transcriptional analysis of dynamic heat-shock response by the hyperthermophilic bacterium *Thermotoga maritima*. Extremophiles.

52. **Reysenbach, A. L., G. S. Wickham, and N. R. Pace.** 1994. Phylogenetic analysis of the hyperthermophilic pink filament community in Octopus Spring, Yellowstone National Park. Appl. Environ. Microbiol. **60:**2113-2119.

53. **Richmond, C. S., J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner.** 1999. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res. **27:**3821-35.

54. **Richmond, C. S., J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner.** 1999. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res. **27:**3821-3835.

55. **Schena, M., D. Shalon, R. W. Davis, and P. O. Brown.** 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270:**467-470.

56. **Schuster, M., C. P. Lostroh, T. Ogi, and E. P. Greenberg.** 2003. Identification, timing, and signal specificity of *Pseudomonas aeruginosa* quorum-controlled genes: a transcriptome analysis. J. Bacteriol. **185:**2066-2079.

57. **Schut, G. J., S. D. Brehm, S. Datta, and M. W. Adams.** 2003. Whole-genome DNA microarray analysis of a hyperthermophile and an archaeon: *Pyrococcus furiosus* grown on carbohydrates or peptides. J. Bacteriol. **185:**3935-3947.

58. **Schut, G. J., J. Zhou, and M. W. Adams.** 2001. DNA microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus*: evidence for anNew type of sulfur-reducing enzyme complex. J. Bacteriol. **183:**7027-7036.

59. **Shockley, K. R., D. E. Ward, S. R. Chhabra, S. B. Conners, C. I. Montero, and R. M. Kelly.** 2003. Heat shock response by the hyperthermophilic archaeon *Pyrococcus furiosus*. Appl. Environ. Microbiol. **69:**2365-2371.

60. **Sibson, D. R. a. S., M. P.** 1997. Increasing the average abundance of low-abundance cDNAs by ordered subdivision of cDNA populations, p. 13-32. *In* I. G. a. A. Cowell, D. A. (ed.), cDNA Library Protocols. Humana Press, New Jersey.

61. **Smulski, D. R., L. X. L. Huang, M. P. McCluskey, M. J. G. Reeve, A. C. Vollmer, M. K. van Dyk, and R. A. LaRossa.** 2001. Combined, functional genomic-biochemical approach to intermediary metabolism: interaction of acivicin, a glutamine amidotransferase inhibitor, with *Escherichia coli* K-12. J. Bacteriol. **183:**3353-3364.

62. **Southern, E., K. Mir, and M. Shchepinov.** 1999. Molecular interactions on microarrays. Nat. Genet. **21:**5-9.

63. **Steil, L., T. Hoffmann, I. Budde, U. Volker, and E. Bremer.** 2003. Genome-wide transcriptional profiling analysis of adaptation of *Bacillus subtilis* to high salinity. J. Bacteriol. **185:**6358-6370.

64. **Stetter, K. O.** 1999. Extremophiles and their adaptation to hot environments. FEBS Lett. **452:**22-25.

65. **Stetter, K. O.** 1998. Hyperthermophiles: isolation, classification, and properties, p. 1-24. *In* K. Horikoshi, and Grant, W. D. (ed.), Extremophiles: Microbial Life in Extreme Environments. Wiley-Liss, Inc., New York.

66. **Stetter, K. O.** 1996. Hyperthermophilic prokaryotes. FEMS Microbiol. Rev. **18:**149-158.

67. **Stintzi, A.** 2003. Gene expression profile of *Campylobacter jejuni* in response to growth temperature variation. J. Bacteriol. **185:**2009-2016.

68. **Thomm, M.** 1996. Archaeal transcription factors and their role in transcription initiation. FEMS Microbiol. Rev. **18:**159-171.

69. **Tschermak, E.** 1900. Über Künstliche Kreuzung bei *Pisum sativum*. Berichte der Deutsche Botanischen Gesellschaft **18:**232-239.

70. **Tucker, D. L., N. Tucker, and T. Conway.** 2002. Gene expression profiling of the pH response in *Escherichia coli* to acetate and propionate. J. Bacteriol. **184:**6551-6558.

71. **Velculescu, V. E., L. Zhang, B. Vogelstein, and K. W. Kinzler.** 1995. Serial analysis of gene expression. Science **270:**484-487.

72. **Wagner, V. E., D. Bushnell, L. Passador, A. I. Brooks, and B. H. Iglewski.** 2003. Microarray analysis of *Pseudomonas aeruginosa* quorum-sensing regulons: effects of growth phase and environment. J. Bacteriol. **185:**2080-2095.

73. **Wang, Q., J. G. Frye, M. McClelland, and R. M. Harshey.** 2004. Gene expression patterns during swarming in *Salmonella typhimurium*: genes specific to surface growth and putative new motility and pathogenicity genes. Mol. Microbiol. **52:**169-187.

74. **Watson, J. D., and F. H. Crick.** 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature **171:**737-738.

75. **Whiteley, M., M. G. Bangera, R. E. Bumgarner, M. R. Parsek, G. M. Teitzel, S. Lory, and E. P. Greenberg.** 2001. Gene expression in *Pseudomonas aeruginosa* biofilms. Nature **413:**860-864.

76. **Xiang, C. C., and Y. Chen.** 2000. cDNA microarray technology and its applications. Biotechnol. Adv. **18:**35-46.

77. **Zamost, B. L., Nielsen, H. K., and Starnes, R. L.** 1991. Thermostable enzymes for industrial applications. J. Indust. Micrbiol. **8:**71-82.

78. **Zhang, J. S., E. L. Duncan, A. C. Chang, and R. R. Reddel.** 1998. Differential display of mRNA. Mol. Biotechnol. **10:**155-165.

79. **Zheng, M., X. Wang, L. J. Templeton, D. R. Smulski, R. A. LaRossa, and G. Storz.** 2001. DNA microarray-mediated transcriptional profiling of the *Escherichia coli* response to hydrogen peroxide. J. Bacteriol. **183:**4562-4570.

80. **Zhu, J., and J. J. Mekalanos.** 2003. Quorum sensing-dependent biofilms enhance colonization in *Vibrio cholerae*. Dev. Cell **5:**647-656.

Table 1. Relevant Prokaryotic DNA Microarray Studies Used for Gene Expression Analysis

Organism	OGT	Microarray Description		Experiment Description	Validation	Ref.
Gram-Positive Bacteria						
Bacillus subtilis	37°C	full	PCR	Sulfur source (sulfate versus methionine)	β-galactosidase activities	(4)
Bacillus subtilis	37°C	full	oligos	Heat shock response	QPCR, promoter comparisons	(25)
Bacillus subtilis	37°C	full	PCR	High salinity	Northern blots and cell motility assays	(63)
Streptococcus pneumoniae	37°C	full	PCR	Wild-type versus mutant mgrA strains	RNAse protection and adherence assays	(26)
Corynebacterium glutamicum	37°C	targeted	PCR	Lysine production and microarray reproducibility	---	(38)
Gram-Negative Bacteria						
Pseudomonas aeruginosa	37°C	full	PCR	Biofilm development	Northern blots and RNAse protection assays	(75)
Pseudomonas aeruginosa	37°C	full	oligos	Quorum Sensing	---	(56)
Pseudomonas aeruginosa	37°C	full	oligos	Quorum Sensing	QPCR	(72)
Thermotoga maritima	80°C	targeted	PCR	Sugar utilization patterns	---	(11)
Thermotoga maritima	80°C	targeted	PCR	Heat shock response	QPCR	(51)
Escherichia coli	37°C	full	PCR	Heat shock response and IPTG induction	spot blots	(54)

Table 1 (cont.)

Escherichia coli	37°C	targeted	PCR	Effect of glucose, acetate, and glycerol levels	metabolic fluxes[a]	(46)
Campylobacter jejuni	42°C	full	PCR	Time-dependent response to temperature increase	QPCR	(67)
Vibrio cholerae	37°C	full	PCR	Wild-type verus quorum sensing-deficient mutants	S1 nuclease protection, CAI-1 production and infant mouse assays	(80)
Salmonella typhimurium	37°C	full	PCR	Swarming	---	(73)
Archaea						
Pyrococcus furiosus	98°C	targeted	PCR	Effect of the present of S^0	---	(58)
Pyrococcus furiosus	98°C	full	PCR	Effect of growth substrate (maltose or peptides)	Enzymatic assays and determination of organic acids	(57)
Pyrococcus furiosus	98°C	targeted	PCR	Heat shock response	Northern blots	(59)

[a]metabolic flux values that were used in this study were taken from values previously published in the literature

Chapter 2: Applications of Genomic Data – Enzyme Discovery and Microbial Genomics

Robert M. Kelly* and Keith R. Shockley

Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

Running Title: Enzyme Discovery and Microbial Genomics

Published in: Microbial Genomics **(The Humana Press, Inc.)**

Claire M. Fraser, Timothy Read and Karen E. Nelson (Editors)

Published in April, 2004

Corresponding author:* **Robert M. Kelly
Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

Phone: (919) 515-6396
Fax: (919) 515-3465
Email: rmkelly@eos.ncsu.ed

Introduction

More than two decades ago, the enzyme industry was focused on a small number of biocatalysts derived from a narrow phylogenetic range of microorganisms. These enzymes were used for a limited set of applications, mostly in large-scale processes for starch processing and as cleaning agents in detergents (1-3). The difficulties associated with discovering new enzymes, producing them on large scale, and integrating them, sometimes imperfectly, into existing bioprocesses constrained the development of this technological sector. Even if a microorganism with physiological features that hinted at the existence of an important biocatalyst could be found, isolating a particular enzyme from among the thousands of similar biomolecules was especially challenging. If the microorganism (wild-type or mutant) had no phenotype that generated copious amounts of the enzyme of interest in large-scale fermentation, the likelihood that this biocatalyst would find its way into a technologically significant use was low. Simply put, before the dawn of industrial biotechnology in the 1980's, commercializing a valuable enzyme embedded in the genotype of an uncharacterized microbial source was the domain of a limited number of industrial practitioners.

But that was then and this is now. Beginning with the first successful commercialization of enzymes produced in recombinant organisms (often involving amplification of gene expression in the native host; 1), the tools of molecular biology have been exploited to completely revise the nature of enzyme discovery (4). Once there were few enzymes considered for many potential applications; now, there is a virtually unlimited number that can be considered based on information encoded in sequenced microbial genomes. The challenge today is to find a technological fit for particular biocatalysts.

Furthermore, recombinant approaches have been used to create engineered versions of specific enzymes that can be customized to the application through site-directed mutagenesis, deoxyribonucleic acid (DNA) shuffling and directed evolution (5, 6). At this point, those focused on biocatalyst technology are seeking answers to questions concerning the diversity of nature's treasure chest of enzymes and whether biological routes can replace or create processes for important chemicals and biochemicals.

Many industrial enzymes have traditionally come from microorganisms isolated from natural environments or from specific niches whereby a biological process mimics the industrial use (4). However, it is clear that conventional methods for enzyme discovery have revealed only a fraction of nature's biocatalytic inventory. Comparative analyses based on small subunit (16S or 18S) ribosomal ribonucleic acid (rRNA) sequences demonstrate that most of the life on Earth is microbial, yet it is estimated that more than 99% of microorganisms observed in nature have not been cultivated by standard techniques (7, 8). Furthermore, fewer than 2% of all microorganisms have been identified, and fewer than that have been closely examined (9). The question arises as to whether the current sampling of microbial enzymes that has been studied represents the larger set. With available microbial genome sequence data, this issue can begin to be investigated.

It is not clear how much microbial biodiversity has been captured in the microbial genomes sequenced to date. Certainly, significant genetic diversity is present in sequenced microbial genomes, even among closely related species or strains within the same species (10). Although pathogenic organisms have been the subjects of most genomic comparison studies to date (11-15), nonpathogenic organisms have been examined as well. For instance, almost one third of the open reading frames (ORFs) in *Bacillus halodurans* did not have a

clear match to those in another organism in its genus, *Bacillus subtilus* (10). Also, the genome of the hyperthermophilic archaeon *Pyrococcus furiosus* (16) is about 10% larger than the genome another organism in this genus, *Pyrococcus horikoshii* (17). Most of the differences can be attributed to additional amino acid biosynthetic pathways and routes for the uptake of carbohydrates, such as cellobiose, maltose, trehalose, laminarin, and chitin (16, 18). The very limited perspective gained from interspecies genome sequence comparisons suggests that, at this point, only a glimpse of nature's biocatalytic repertoire has been seen.

Clearly, the gene-by-gene, protein-by-protein approach to enzyme discovery is being replaced by the use of new methodologies and information arising from advances in the molecular biology and genomics sciences (9, 19, 20). When complete microbial genome sequences appeared in the mid-1990's, the first look into the complete enzymatic inventory within microbial species was possible. Today, about 60 microbial genomes have been sequenced and at least 184 genome sequences are in progress (21). For the first time, it has been possible to compare indirect evidence for the presence of specific biocatalysts in an organism against its own genetic blueprint.

Much uncertainty still exists, given that half or more of the ORFs in microbial genomes were initially not coupled to specific function. For instance, even in the much-studied *Escherichia coli*, with 4288 annotated protein-coding genes, 38% initially had no attributed function (22). Also, of the 2,977 proteins predicted to be encoded by the extremely thermoacidophilic crenarchaeon *Sulfolobus solfataricus* P2, about one-third have no detectable homologs in other sequenced genomes (23). Over 50% of the ORFs discovered in the *P. horikoshii* genome have functions that have not been identified through database similarity searches (24). The correct annotation of ORFs in microbial genomes remains an

ongoing effort that currently employs a diverse set of tools including *in vivo*, *in vitro* and *in silico* approaches. As has been the case to date, the annotation process has probably raised as many new questions as it has provided answers. Nonetheless, the technologically-relevant enzyme inventory for particular microorganisms is becoming more clear.

Armed with the prospect of using PCR to amplify a gene of interest from genomic DNA for subsequent cloning and overexpression in a suitable host, each microbial genome can potentially provide many biocatalysts that can be examined for technological importance. Selection of candidates for further characterization can be facilitated through the use of a variety of bioinformatics tools (see Table 1). If several candidates are identified, the list of those to be produced and characterized biochemically can be shortened considerably by examining structural and catalytic traits *in silico*, preferably in conjunction with amino acid sequence-based classification schemes, such as those developed for glycosyl hydrolases (25).

Mining hyperthermophile genomes for useful biocatalysts

For several reasons, initial efforts for sequencing microbial genomes focused on microorganisms that typically inhabit biologically-hostile environments (26). Hyperthermophiles, extremophilic microorganisms that belong to the domains Archaea and Bacteria, have optimal growth temperatures of 80°C or higher. Stemming from their evolutionary placement, small genome size, and potential importance as a source of stable biocatalysts, a number of genome sequences for these microorganisms have been reported (see Table 2).

Although a myriad of recombinant techniques now exist for improving enzyme characteristics (see below), these are best applied to a biocatalyst with natural properties

close to the final optimal form. One particularly sought-after trait in industrial biocatalysts is stability to heat, an intrinsic feature of enzymes from hyperthermophiles. The interest in thermostable biocatalysts arises from the fact that most industrial processes still utilize enzymes from mesophilic microorganisms, even though many reactions are conducted at elevated temperatures (4, 27-29). At higher temperatures, lower enzyme concentrations and higher conversions may result from decreased viscosity and larger diffusion coefficients of organic compounds (27, 30, 31). Elevated temperatures also can lead to higher substrate bioavailability, along with reduced risks of biological contamination (26, 27, 31). As an added benefit, enzymes from hyperthermophilic organisms are often relatively resistant to chemical denaturants, such as detergents, chaotropic agents, and organic solvents, making them useful as industrial biocatalysts (26, 31-33). Thus, if stability is important, an already hyperthermophilic enzyme can be modified through recombinant approaches to improve catalytic traits. Given the number of hyperthermophile genome sequences completed and available, there is a good possibility of finding a thermostable enzyme with biocatalytic properties that either match specific needs or can be modified to meet specific requirements.

Examining the biocatalyst inventory in hyperthermophile genomes

Most often, ORFs in genomic sequence data are annotated through full sequence alignment (e.g., Basic Local Alignment Search Tool [BLAST]; 34) to those found in databases such as GenBank. Similar approaches can be used to determine the inventory of specific enzymes among selected organisms, assisted by specialized databases (e.g., ref. 35) and handbooks (e.g., ref. 36). For example, Table 3 shows the inferred protease inventory from publicly available genome sequences for hyperthermophiles, based on all known

(isolated and characterized biochemically) or putative (inferred from bioinformatics tools, such as those shown in Table 1) proteases in *P. furiosus* (37). The protease homologs in Table 1 were defined based on more than 30% amino acid sequence identity over more than 50% of the sequence; this criterion is arbitrary and can be relaxed or made more stringent. Table 3 illustrates the biodiversity of proteases within a given organism and between organisms. As might be expected, the three pyrococci examined share many similar protease-encoding genes, although they also clearly exhibit differences. In some cases, there are no homologs to the *P. furiosus* proteases among the hyperthermophiles listed, while in other cases homologs appear to exist but can differ significantly in molecular mass. To illustrate, a putative protease (with a signal peptide; PF1905), three aminopeptidases (with signal peptides; PF2059, PF2063, and PF2065), an intracellular putative bacteriocin/protease (PF1191), and a transmembrane protease, pyrolysin (PF0287), appear to be unique to *P. furiosus*; an intracellular o-sialoglycoprotein endopeptidase (PF0172), a proline dipeptidase (PF1343) and a methionine dipeptidase (PF0541) are ubiquitous in the hyperthermophile genome sequences examined (Table 3).

Some of the first enzymes studied from hyperthermophiles were glycosyl hydrolases (38, 39). These attracted attention because of their industrial significance in the starch processing industries and their physiological importance for heterotrophic hyperthermophiles growing on glycan-based media (40, 41). Hyperthermophile genome sequences revealed significant differences among hyperthermophiles in their available enzyme inventory for carbohydrate hydrolysis (42). Although the genome sequence for the hyperthermophilic archaeon *P. furiosus* (43) shows the presence of a range of glucan-degrading enzymes (Table 4), the genome sequence of *Archaeglobus fulgidus* (44), a hyperthermophilic archaeon, is

apparently devoid of such enzymes. Furthermore, even among three pyrococci, glycosyl hydrolase content varies, especially with respect to enzymes capable of degrading laminarin (40) and chitin (Gao and Kelly, unpublished data). In fact, *P. furiosus* appears to contain a chitin utilization pathway that includes a chitin deactylase and glucoaminidase not evident from initial genome annotations (Gao and Kelly, unpublished data). Also, the *Methanococcus jannaschii* genome (45) revealed the presence of several glycosidases, one of which (MJ1610) is a family 15 glucoamylase. Given the small sampling of hyperthermophilic genomes, it is difficult to assess the diversity of enzymes such as glycosidases, although one should expect many surprises as these organisms are probed further.

At one level, genome sequence annotation provides a glimpse into the actual and putative enzymatic inventory of a specific microorganism. However, genome sequence annotation itself presents significant challenges for biocatalyst identification and the results can be misleading at times (20, 46). For instance, two putative enzymes that share a similar substrate-binding domain may have dissimilar catalytic domains, although a sequence alignment tool may classify both as related, and reflect this in the annotation. Furthermore, when simple homology searches based on full sequence length are the basis for annotation, these may not detect less obvious relationships found in enzyme super families and will not recognize non-orthogolous genes carrying the same function (4, 47, 48). Another problem occurs when incorrect functions are assigned to specific ORFs in databases and this assignment propagates in subsequent sequences that are reported. Analysis based on simple amino acid sequence homology alone is not powerful enough to confirm the absence of an enzyme from a genome sequence, or rule out the possibility that an enzyme has multiple functions. Incomplete understanding of cellular metabolism also creates problems. For

instance, some of the genes encoding enzymes that are used in microbial tryptophan biosynthesis pathways were missing from the genome of *P. horikoshii* (24). Although this organism requires tryptophan for cell viability and growth, it is not clear whether it is auxotrophic for this amino acid or whether a complete synthetic pathway exists but contains unidentified elements.

When BLAST searches alone are not sufficient to elucidate gene function, other bioinformatic approaches may complement the analyses. For example, enzymatic resolution of racemic mixtures of 2-aryl propionic esters, such as those used for non-steroidal anti-infammatory drugs (49), has been reported using esterases and lipases from mesophilic sources. BLAST searches, in conjunction with protein structure-based motif analysis (TOPITS; 50), identified a carboxylesterase in the genome of *S. solfataricus* P1 (SsoEST1) (51). This enzyme proved to be more effective for the resolution of Naproxen methyl esters than other mesophilic candidates, despite the fact that temperatures more than 50°C below this enzyme's optimum were used (52). While BLAST searches alone turned up other thermostable esterases/lipases posibilities, the combination of several bioinformatics tools led to the most promising candidate.

Searching databases, such as PROSITE (53), for short sequence patterns or motifs to identify functional domains in predicted proteins can eliminate some of the problems associated with matching entire sequences (54). For instance, a database termed IDENTIFY (55), may be able to identify a protein superfamily, even when BLAST results are not suggestive. A total of 833 ORFs in the yeast genome had not been assigned a function when the genome was first published, but 172 of the unknown proteins were subsequently assigned putative functions based on the IDENTIFY algorithm (55).

Other approaches have also been used. Threading or *ab initio* folding methods allow the prediction of tertiary structure from sequence information, and can be further screened by descriptors of protein active sites called *fuzzy functional forms* (47, 54). This approach helped to identify the function of two proteins in the glutaredon/thioreoxin disulphide oxidoreductase family in the yeast genome whose functions could not be predicted by BLAST searches or local sequence alignment algorithms (54). Another approach, based on combining methods of prediction and experimental data, examined correlated evolution, correlated messenger RNA expression patterns, and patterns of domain fusion among the 6,217 proteins of the yeast *Saccharomyces cerevisiae* to assign general functions to more than half of the 2,557 previously uncharacterized yeast proteins (56). Both methods demonstrate that general protein functions encoded by open reading frames in DNA sequences can be assigned based on functional relatedness criteria distinct from amino acid sequence similarity alone.

Intergenomic comparisons among microbial genomes can also be used to find promising biocatalysts. For example, conservation of gene clusters across a wide range of organisms can help to determine candidates for homologous function or indicate the presence of an essential role. The malaria parasite *Plasmodium falciparum* uses an essential isoprenoid biosynthesis pathway commonly found in plants but not in mammals, which led to the discovery of a herbicide that targets this pathway as a specific anti-malarial agent (48, 57). Although lateral gene transfer has complicated single gene studies of evolutionary relatedness, full genome sequence information can be used to identify horizontally transferred genes more rapidly by looking for regional differences, such as base composition in bacterial chromosomes (48).

Functional genomics and enzyme discovery

While genomic information pertaining to identifiable ORFs can provide a useful blueprint of an organism's repertoire of genes, the challenge remains to discover how an organism uses its genetic information to accomplish biological tasks. From this kind of insight, purposeful uses of specific enzymes can be projected. Genetic regulation has been described as the process by which a cell decides whether a gene is active or inactive (58, 59). Active genes are those that are being transcribed, whose transcripts are being translated, and whose enzymatic products are actively performing their function. Functional genomics includes both transcriptional analyses and proteomics approaches, making use of gene expression data, systematic mass mutagenesis, and protein interaction maps in order to try to elucidate functions of genes (47, 60). In the bacteria and the eukarya (and presumably the archaea as well), the majority of gene regulation is most often controlled at the level of transcription initiation (48, 61, 62). Therefore, an understanding of the mechanisms that regulate the initiation of gene transcription is a good way to discover enzymes of interest and is essential to the knowledge of the underlying biological processes.

As mentioned, genome sequence comparison is a principal first step toward elucidating the metabolic role of a given protein encoded by genomic sequence information, but results from bioinformatic predictions need to be confirmed through both transcriptional and biochemical analyses. For example, *Thermotoga maritima*, a hyperthermophilic bacterium capable of growth on a spectrum of α- and β-linked glycosides (63), produces several glycosyl hydrolases including an endoglucanase (Cel5A) and a mannanase (Man5). When compared against proteins found in the GenBank database through BLAST searching (42), Man5 was most similar (46% identity at the amino acid sequence level) to a β-

mannanase (ManF) from *Bacillus stearothermophilus*, while Cel5A showed highest similarity (38% amino acid sequence identity) to a family 5 endoglucanase (CelD) from *Clostridium celluloyticum*. Northern blotting and cDNA microarray experiments demonstrated that *man5* was induced when *T. maritima* was grown on carob galactomannan, konjac glucomannan and to a lesser extent on carboxymethyl cellulose. Surprisingly, *cel5A* was induced only on konjac glucomannan.

To investigate this unexpected result further, the activities of recombinant forms of Man5 and Cel5A were tested against a variety of polysaccharide substrates. Man5 was active only on mannose-based polysaccharides; Cel5A was active against glucans, xylans and mannans. Remarkably, the activity of Cel5A was comparable to that of Man5 against galactomannan and higher than Man5 against glucomannan, a finding much different from what was predicted by genome sequence comparison alone. Further investigation of the two enzymes revealed that Man5 contained a signal peptide; Cel5A did not. Taken together, these results indicate that the primary physiological role of Cel5A (and a related enzyme Cel5B) is to break down glucomannan oligosaccharides that are transported into the cell following extracellular hydrolysis by exported glycosidases, such as Man5, even though the genes encoding Cel5A and Man5 are not proximal in the genome (see Figure 1). The functional assignment for Cel5A is based on biochemical properties of the enzyme in conjunction with gene regulation patterns for the native organism growing on various substrates.

Biological function of an enzyme encoded on a sequenced genome can sometimes be determined using database comparisons of sequence data with previously characterized proteins or bioinformatic tools. These include the Kyoto Encyclopedia of Gene and Genomes (or KEGG) (64) and EcoCyc (65), both of which present genomic information in a

comprehensive form based on biological pathways and molecular assemblies. However, it is also useful to have functional data directly from expression analysis studies through the identification and quantification of differentially expressed transcripts (DETs) between two or more biologically varying samples.

Numerous advances have been made that allow the identification and quantification of differentially expressed genes, especially cDNA microarrays (66-73), which were first used to simultaneously monitor the expression of 45 different *Arabidopsis* genes (74). Since the debut of this technology, DNA microarrays have been used to study relative wide-scale gene expression patterns in many different kinds of organisms, including bacteria (66-73), archaea (75), yeast (76-78), plants (79, 80), fruit flies (81), mice (82), and human beings (83-85). Microarrays provide a mechanism to identify differentially expressed genes in microorganisms growing on a specific substrate, the modification of which is of technological importance. For example, the genes encoding Cel5A and Man5 in *T. maritima* were indeed found to be co-regulated during growth on mannan-based compounds using targeted cDNA microarrays (86). As microarray technology improves and becomes less expensive to use, one can envision using environmental arrays to follow gene expression patterns in microbial consortia as a means of biocatalyst discovery.

Biocatalyst design by genome scanning, functional screening and improvement

High throughput screening techniques are improving every year (87, 88), such that recombinant approaches for biocatalyst improvement can produce enzymes for specific applications with optimal properties. Directed molecular evolution and related methods can be used to generate useful variants of existing enzymes, such that the resulting biocatalyst

works better than the natural one. This approach consists of multiple rounds of mutagenesis and screening, followed by amplification of selected variants (4, 5, 89). As genome annotation becomes more sophisticated, so too will selection processes for biomolecules that can serve as strategic starting points for evolution techniques.

Directed molecular evolution was used to create an ampicillin-resistance activity from a functionally unrelated DNA fragment in *P. furiosus* (89). The resulting mutant enzyme conferred resistance to other drugs that target bacterial cell wall synthesis as well, although the mechanism of action is unclear. Because *P. furiosus* is a hyperthermophilic archaeon, it does not contain peptidoglycan in its cell wall and is not susceptible to common antibiotics that are directed against the synthesis of cell walls, including β-lactam antibiotics like ampicillin. Nevertheless, an expression library of *P. furiosus* DNA fragments was screened for a gene that created ampicillin-resistance activity (amp^R) in *E. coli*, which corresponded to a 1.2-kb fragment (including an ORF encoding a 226 amino acid protein) (89). This DNA fragment was subjected to 50 rounds of directed evolution, in which mutations and DNA recombinations were randomly introduced. The resulting DNA fragments contained two genetic regions that had coevolved during the course of the experiment; one region was essential for the ampicillin resistance activity while the other was able to enhance the activity. This experiment illustrates the potential utility of choosing genes from genome sequences that can be evolved to function in applications unrelated to their natural roles.

Microbial genomics: Future directions for enzyme discovery

It is still too early to tell how effective genomics-based enzyme discovery approaches will be. Beyond identifying homologs to enzymes of interest through bioinformatics tools,

there is great interest in novel physiological systems that give rise to unique enzymes or pathways of technological importance. These can be inferred from differential expression experiments in which whole genome or targeted microarrays are used to follow genetic response to environmental and nutritional changes. For example, *P. furiosus*, like other hyperthermophiles, lacks a phosphotransferase system for carbohydrate uptake, instead relying on adenosine-triphosphate binding cassette (ABC) transporters (see Figure 2). Glycosidase-transporter couplings can be used to track cellular response to various carbohydrates, thereby providing hints to yet to be annotated genes encoding enzymes capable of hydrolyzing substrates of interest. Thus, new enzymes involved in various stages of polysaccharide hydrolysis can be identified from differential expression analysis on specific carbohydrates. Similar approaches for other types of enzymes can be used if sufficient insight into the organism's physiological patterns is available.

Much has changed in the past two decades with respect to enzyme discovery. The arrival of microbial genomics will no doubt stimulate creative approaches to finding biocatalysts that until this point have been hidden within certain microbial genotypes. By combining newly developed methods for high throughput screening, directed evolution, and biocatalyst production with bioinformatics tools, microbial genomes can be fully utilized for significant technological advances related to important biotransformations.

Acknowledgments

This work was supported in part by grants from the Biotechnology Program, National Science Foundation and the Energy Biosciences Program, U.S. Department of Energy.

References

1. Dordick, J.S. (1991) The General Uses of Biocatalysts. In *Biocatalysts for Industry* (J. S. Dordick, ed.), pp. 1-19. Plenum Press, New York.
2. Neidleman, S.L. (1991) Historical Perspective on the Industrial Uses of Biocatalysts. In *Biocatalysts for Industry* (J. S. Dordick, ed.), pp. 21-33. Plenum Press, New York.
3. Uhlig, H. (1998) *Industrial Enzymes and Their Applications*. Trans. Elfriede M. Linsmaier-Bednar, John Wiley & Sons, Inc., New York.
4. Marrs, B., Delagrave, S., Murphy, D. (1999) Novel approaches for discovering industrial enzymes. Curr. Opin. Microbiol. **2**:241-245.
5. Arnold, F.H., Volkov, A.A. (1999) Directed evolution of biocatalysts. Curr. Opin. Chem. Biol. **3**, 54-59.
6. Crameri, A., Raillard, S.A., Bermudez, E., Stemmer, W.P. (1998) DNA shuffling of a family of genes from diverse species accelerates directed evolution. Nature **391**:288-291.
7. Aaman, R.I., Ludwig, W., Schleifer, K.-H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. Microbiol. Rev. **59**:143-169.
8. Hugenholtz, P., Goebel, B.M., Pace, N.R. (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J. Bacteriol. **180**:4765-4774.
9. Bull, A.T., Ward, A.C., Goodfellow, M. (2000) Search and discovery strategies for biotechnology: the paradigm shift. Microbiol. Mol. Biol. Rev. **64**:573-606.

10. Boucher, Y., Nesbo, C.L., Doolittle, W.F. (2001) Microbial genomes: dealing with diversity. Curr. Opin. Microbiol. **4**:285-289.

11. Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R.L., Zhao, Q., Koonin, E.V., Davis, R.W. (1998) Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*. Science **282**:754-759.

12. Alm, R.A., Ling, L.S., Moir, D.T., King, B.L., Brown, E.D., Doig, P.C., Smith, D.R., Noonan, B., Guild, B.C., deJonge, B.L., Carmel, G., Tummino, P.J., Caruso, A., Uria-Nickelsen, M., Mills, D.M., Ives, C., Gibson, R., Merberg, D., Mills, S.D., Jiang, Q., Taylor, D.E., Vovis, G.F., Trust, T.J. (1999) Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. Nature **397**:176-180.

13. Alm, R.A., Trust, T.J. (1999) Analysis of the genetic diversity of *Helicobacter pylori*: the tale of two genomes. J. Mol. Med. **77**:834-846.

14. Kalman, S., Mitchell, W., Marathe, R., Lammel, C., Fan, J., Hyman, R.W., Olinger, L., Grimwood, J., Davis, R.W., Stephens, R.S. (1999) Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*. Nat. Genet. **21**:385-389.

15. Read, T.D., Brunham, R.C., Shen, C., Gill, S.R., Heidelberg, J.F., White, O., Hickey, E.K., Peterson, J., Utterback, T., Berry, K., Bass, S., Linher, K., Weidman, J., Khouri, H., Craven, B., Bowman, C., Dodson, R., Gwinn, M., Nelson, W., DeBoy, R., Kolonay, J., McClarty, G., Salzberg, S.L., Eisen, J., Fraser, C.M. (2000) Genome sequences of *Chlamydia trachomatis MoPn* and *Chlamydia pneumoniae AR39*. Nucleic Acids Res. **28**:1397-1406.

16. Robb, F.T., Maeder, D.L., Brown, J.R., DiRuggiero, J., Stump, M.D., Yeh, R.K., Weiss, R.B., Dunn, D.M. (2001) Genomic sequence of hyperthermophile*, Pyrococcus furiosus*: Implications for physiology and enzymology. Methods Enzymol. **330**:134-157.

17. Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Kikuchi, H. (1998) Complete sequence and gene organization of the genome of a hyper- thermophilic archaebacterium, *Pyrococcus horikoshii* OT3 (supplement). DNA Res. **5**:147-155.

18. Lecompte, O., Ripp, R., Puzos-Barbe, V., Duprat, S., Heilig, R., Dietrich, J., Thierry, J.C., Poch, O. (2001) Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. *Genome Res.* **11**:981-993.

19. Dean, P.M., Zanders, E.D., Bailey, D.S. (2001) Industrial-scale, genomics-based drug design and discovery. Trends Biotechnol. **19**:288-292.

20. Tang, C.M., Moxon, E.R. (2001) The impact of microbial genomics on antimicrobial drug development. *Annu. Rev. Genomics Hum. Genet.* **2**:259-269.

21. The Institute for Genomic Research. Completed genomes can be found at http://www.tigr.org/tdb/mdb/mdbcomplete.html; genomes in progress can be found at http://www.tigr.org/tdb/mdb/mdbinprogress.html.

22. Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., ColladoVides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W.,

Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453-1462.

23. She, Q., Singh, R.K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M.J., Chan-Weiher, C.C.Y., Clausen, I.G., Curtis, B.A., De Moors, A., Erauso, G., Fletcher, C., Gordon, P.M.K., Heikamp-de Jong, I., Jeffries, A.C., Kozera, C.J., Medina, N., Peng, X., Thi-Ngoc, H.P., Redder, P., Schenk, M.E., Theriault, C., Tolstrup, N., Charlebois, R.L., Doolittle, W.F., Duguet, M., Gaasterland, T., Garrett, R.A., Ragan, M.A., Sensen, C.W., Van der Oost, J. (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc. Natl. Acad. Sci. USA. **98**:7835-7840.

24. Kawarabayasi, Y. (2001) Genome of *Pyrococcus horikoshii* OT3. Methods Enzymol. **330**:124-134.

25. Henrissat, B., Teeri, T.T., Warren, R.A. (1998) A scheme for designating enzymes that hydrolyse the polysaccharides in the cell walls of plants. FEBS Lett. **425**:352-354.

26. Adams, M.W., Perler, F.B., Kelly, R.M. (1995) Extremozymes: expanding the limits of biocatalysis. Biotechnol. **13**:662-668.

27. Zamost, B.L., Nielsen, H.K., Starnes, R.L. (1991) Thermostable enzymes for industrial applications. J. Indust. Microbiol. **8**:71-82.

28. Stetter, K.O. (1998) Hyperthermophiles: isolation, classification, and properties. In *Extremophiles: Microbial Life in Extreme Environments* (K. Horikoshi and W. D. Grant, eds.), pp. 1-24. Wiley-Liss, Inc., New York.

29. Demirjian, D.C., Moris-Varas, F., Cassidy, C.S. (2001) Enzymes from extremophiles. Curr. Opin. Chem. Biol. **5**:144-151.

30. Kalisz, M.H. (1988) Microbial proteinases. Adv. Biochem. Eng. Biotechnol. **36**:17-55.

31. Niehaus, F., Bertoldo, C., Kahler, M., Antranikian, G. (1999) Extremophiles as a source of novel enzymes for industrial application. Appl. Microbiol. Biotechnol. **51**:711-729.

32. von der Osten, C., Branner, S., Hastrup, S., Hedegaard, L., Rasmussen, M.D., Bisgard-Frantzen, H., Carlsen, S., Mikkelsen, J.M. (1993) Protein engineering of subtilisins to improve stability in detergent formulations. J. Biotechnol. **28**:55-68.

33. Cowan, D.A. (1995) Protein stability at high temperatures. Essays Biochem. **29**:193-207.

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

35. Barrett, A.J., Rawlings, N.D., O'Brien, E.A. (2001) The MEROPS database as a protease information system. J. Struct. Biol. **134**:95-102.

36. Barrett, A.J., Rawlings, N.D., Woessner, J.F. (1998) Handbook of Proteolytic Enzymes. Academic Press, London.

37. Ward, D.E., Shockley, K.R., Chang, L.S., Levy, R.D., Michel, J.K., Conners, S.B., Kelly, R.M. (2002) Proteolysis in hyperthermophilic microorganisms. Archaea **1**:63-74.

38. Costantino, H.R., Brown, S.H., Kelly, R.M. (1990) Purification and characterization of an alpha-glucosidase from a hyperthermophilic archaebacterium, *Pyrococcus*

furiosus, exhibiting a temperature optimum of 105 to 115 degrees C. J. Bacteriol. **172**:3654-3660.

39. Brown, S.H. (1992) Saccharidases from High-Temperature Bacteria: Physiological and Enzymological Studies. In *Chemical Engineering*, pp. 216. The Johns Hopkins University, Baltimore.

40. Bauer, M.W., Driskill, L.E., Kelly, R.M. (1998) Glycosyl hydrolases from hyperthermophilic microorganisms. Curr. Opin. Biotechnol. **9**:141-145.

41. Driskill, L.E., Kusy, K., Bauer, M.W., Kelly, R.M. (1999) Relationship between glycosyl hydrolase inventory and growth physiology of the hyperthermophile *Pyrococcus furiosus* on carbohydrate-based media. Appl. Environ. Microbiol. **65**:893-897.

42. Chhabra, S.R., Shockley, K.R., Ward, D.E., Kelly, R.M. (2002) Regulation of endo-acting glycosyl hydrolases in the hyperthermophilic bacterium *Thermotoga maritima* grown on glucan- and mannan-based polysaccharides. Appl. Environ. Microbiol. **68**:545-554.

43. Weiss, R.B. (2002) Direct Submission: *Pyrococcus furiosus* genomic sequence. Human Genetics, University of Utah, 20 South 2030 East, Salt Lake City, UT 84112.

44. Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., Richardson, D.L., Kerlavage, A.R., Graham, D.E., Kyrpides, N.C., Fleischmann, R.D., Quackenbush, J., Lee, N.H., Sutton, G.G., Gill, S., Kirkness, E.F., Dougherty, B.A., McKenney, K., Adams, M.D., Loftus, B., Peterson, S., Reich, C.I., McNeil, L.K., Badger, J.H., Glodek, A., Zhou, L.X., Overbeek, R., Gocayne, J.D., Weidman, J.F., McDonald, L., Utterback, T.,

Cotton, M.D., Spriggs, T., Artiach, P., Kaine, B.P., Sykes, S.M., Sadow, P.W., Dandrea, K.P., Bowman, C., Fujii, C., Garland, S.A., Mason, T.M., Olsen, G.J., Fraser, C.M., Smith, H.O., Woese, C.R., Venter, J.C. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. Nature **390**:364-370.

45. Bult, C.J., White, O., Olsen, G.J., Zhou, L.X., Fleischmann, R.D., Sutton, G.G., Blake, J.A., FitzGerald, L.M., Clayton, R.A., Gocayne, J.D., Kerlavage, A.R., Dougherty, B.A., Tomb, J.F., Adams, M.D., Reich, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S.M., Weidman, J.F., Fuhrmann, J.L., Nguyen, D., Utterback, T.R., Kelley, J.M., Peterson, J.D., Sadow, P.W., Hanna, M.C., Cotton, M.D., Roberts, K.M., Hurst, M.A., Kaine, B.P., Borodovsky, M., Klenk, H.P., Fraser, C.M., Smith, H.O., Woese, C.R., Venter, J.C. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273**:1058-1073.

46. Pennisi, E. (1999) Keeping genome databases clean and up to date. Science **286**:447-450.

47. Pallen, M.J. (1999) Microbial genomes. Mol. Microbiol. **32**:907-912.

48. Sassetti, C., Rubin, E.J. (2002) Genomic analyses of microbial virulence. Curr. Opin. Microbiol. **5**:27-32.

49. Sehgal, D. (2002) Efffect of Reaction Environment on Biocatalysis and Enantioselectivity of Hyperthermophilic Esterases. In *Chemical Engineering*, pp. 129. North Carolina State University, Raleigh.

50. Rost, B. (1995) TOPITS: threading one-dimensional predictions into three-dimensional structures. Proc. Int. Conf. Intell. Syst. Mol. Biol. **3**:314-321.

51. Sehgal, A.C., Callen, W., Mathur, E.J., Short, J.M., Kelly, R.M. (2001) Carboxylesterase from *Sulfolobus solfataricus* P1. Methods Enzymol. **330**:461-471.

52. Sehgal, A.C., Kelly, R.M. (2002) Enantiomeric resolution of 2-aryl propionic esters with hyperthermophilic and mesophilic esterases: contrasting thermodynamic mechanisms. J. Am. Chem. Soc. **124**:8190-8191.

53. Hofmann, K., Bucher, P., Falquet, L., Bairoch, A. (1999) The PROSITE database, its status in 1999. Nucleic Acids Res. **27**:215-219.

54. Fetrow, J.S., Skolnick, J. (1998) Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to glutaredoxins/thioredoxins and T1 ribonucleases. J. Mol. Biol. **281**:949-968.

55. Nevill-Manning, C.G., Wu, T.D., Brutlag, D.L. (1998) Highly specific protein sequence motifs for genome analysis. Proc. Natl. Acad. Sci. USA **95**:5865-5871.

56. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function. Nature **402**:83-86.

57. Jomaa, H., Wiesner, J., Sanderbrand, S., Altincicek, B., Weidemeyer, C., Hintz, M., Turbachova, I., Eberl, M., Zeidler, J., Lichtenthaler, H.K., Soldati, D., Beck, E. (1999) Inhibitors of the nonmevalonate pathway of isoprenoid biosynthesis as antimalarial drugs. Science **285**:1573-1576.

58. Kornberg, R.D. (2000) Eukaryotic transcriptional control. Trends. Biochem. Sci. **24**:M46-M49.

59. Klug, W.S., Cummings, M.R. (2000) *Genetics*, Prentice-Hall, Inc., Upper Saddle River.

60. Kotra, L.P., Vakulenko, S., Mobashery, S. (2000) From genes to sequences to antibiotics: prospects for future developments from microbial genomics. Microbes Infect. **2**:651-658.

61. Neidhardt, F.C., Ingraham, J.L., Schaechter, M. (1990) *Physiology of the bacterial cell: A molecular approach*, Sinauer Associates, Inc., Sunderland, MA.

62. Thomm, M. (1996) Archaeal transcription factors and their role in transcription initiation. FEMS Microbiol. Rev. **18**:159-171.

63. Huber, R., Langworthy, T.A., Konig, H., Thomm, M., Woese, C.R., Sleytr, U.B., Stetter, K.O. (1986) *Thermotoga maritima* sp. nov. represent a new genus of unique extremely thermophilic eubacteria growing up to 90^{o}C. Arch. Microbiol. **144**:324-333.

64. Kanehisa, M., Goto, S., Kawashima, S., Nakaya, A. (2002) The KEGG databases at GenomeNet. Nucleic Acids Res. **30**:42-46.

65. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C., Gama-Castro, S. (2002) The EcoCyc Database. Nucleic Acids Res. **30**:56-58.

66. Richmond, C.S., Glasner, J.D., Mau, R., Jin, H., Blattner, F.R. (1999) Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res. **27**:3821-3835.

67. Tao, H., Bausch, C., Richmond, C., Blattner, F.R., Conway, T. (1999) Functional genomics: expression analysis of *Escherichia coli* growing on minimal and rich media. J. Bacteriol. **181**:6425-6440.

68. Helmann, J.D., Wu, M.F., Kobel, P.A., Gamo, F.J., Wilson, M., Morshedi, M.M., Navre, M., Paddon, C. (2001) Global transcriptional response of *Bacillus subtilis* to heat shock. J. Bacteriol. **183**:7318-7328.

69. Merrell, D.S., Butler, S.M., Qadri, F., Dolganov, N.A., Alam, A., Cohen, M.B., Calderwood, S.B., Schoolnik, G.K., Camilli, A. (2002) Host-induced epidemic spread of the cholera bacterium. Nature **417**:642-645.

70. Oh, M.K., Liao, J.C. (2000) DNA microarray detection of metabolic responses to protein overproduction in *Escherichia coli*. Metab. Eng. **2**:201-209.

71. Oh, M.K., Liao, J.C. (2000) Gene expression profiling by DNA microarrays and metabolic fluxes in *Escherichia coli*. Biotechnol. Prog. **16**:278-286.

72. Loos, A., Glanemann, C., Willis, L.B., O'Brien, X.M., Lessard, P.A., Gerstmeir, R., Guillouet, S., Sinskey, A.J. (2001) Development and validation of *Corynebacterium* DNA microarrays. Appl. Environ. Microbiol. **67**:2310-2318.

73. Oh, M.K., Rohlin, L., Kao, K.C., Liao, J.C. (2002) Global expression profiling of acetate-grown *Escherichia coli*. J. Biol. Chem. **277**:13175-13183.

74. Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science **270**:467-470.

75. Schut, G.J., Zhou, J.Z., Adams, M.W.W. (2001) DNA microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus*: Evidence for a new type of sulfur-reducing enzyme complex. J. Bacteriol. **183**:7027-7036.

76. ter Linde, J.J., Liang, H., Davis, R.W., Steensma, H.Y., van Dijken, J.P., Pronk, J.T. (1999) Genome-wide transcriptional analysis of aerobic and anaerobic chemostat cultures of *Saccharomyces cerevisiae*. J. Bacteriol. **181**:7409-7413.

77. Lashkari, D.A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., Davis, R.W. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc. Natl. Acad. Sci. USA **94**:13057-13062.

78. Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburtty, K., Simon, J., Bard, M., Friend, S.H. (2000) Functional discovery via a compendium of expression profiles. Cell **102**:109-126.

79. Desprez, T., Amselem, J., Caboche, M., Hofte, H. (1998) Differential gene expression in *Arabidopsis* monitored using cDNA arrays. Plant J. **14**:643-652.

80. Kawasaki, S., Borchert, C., Deyholos, M., Wang, H., Brazille, S., Kawai, K., Galbraith, D., Bohnert, H.J. (2001) Gene expression profiles during the initial phase of salt stress in rice. Plant Cell **13**:889-905.

81. Jin, W., Riley, R.M., Wolfinger, R.D., White, K.P., Passador-Gurgel, G., Gibson, G. (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Genet. **29**:389-395.

82. Kaminski, N., Allard, J.D., Pittet, J.F., Zuo, F., Griffiths, M.J., Morris, D., Huang, X., Sheppard, D., Heller, R.A. (2000) Global analysis of gene expression in pulmonary fibrosis reveals distinct programs regulating lung inflammation and fibrosis. Proc. Natl. Acad. Sci. USA **97**:1778-1783.

83. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., Davis, R.W. (1996) Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. Proc. Natl. Acad. Sci. USA **93**:10614-10619.

84. Alizadeh, A.A., Eisen, M.B., Davis, R.E., Ma, C., Lossos, I.S., Rosenwald, A., Boldrick, J.C., Sabet, H., Tran, T., Yu, X., Powell, J.I., Yang, L., Marti, G.E., Moore, T., Hudson, J., Jr., Lu, L., Lewis, D.B., Tibshirani, R., Sherlock, G., Chan, W.C., Greiner, T.C., Weisenburger, D.D., Armitage, J.O., Warnke, R., Staudt, L.M., et al. (2000) Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature **403**:503-511.

85. Coller, H.A., Grandori, C., Tamayo, P., Colbert, T., Lander, E.S., Eisenman, R.N., Golub, T.R. (2000) Expression analysis with oligonucleotide microarrays reveals that MYC regulates genes involved in growth, cell cycle, signaling, and adhesion. Proc. Natl. Acad. Sci. USA **97**:3260-3265.

86. Chhabra, S.R., Shockley, K.R., Conners, S.B., Scott, K., Wolfinger, R.D., Kelly, R.M. (2003) Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima*. J. Biol. Chem. **278**, 7740-7752.

87. Demirjian, D.C., Shah, P.C., Moris-Varas, F. (1999) Screening for novel enzymes. Top. Curr. Chem. **200**:1-29.

88. Dautin, N., Karimova, G., Ullmann, A., Ladant, D. (2000) Sensitive genetic screen for protease activity based on a cyclic AMP signaling cascade in *Escherichia coli*. J. Bacteriol. **182**:7060-7066.

89. Yano, T., Kagamiyama, H. (2001) Directed evolution of ampicillin-resistant activity from a functionally unrelated DNA fragment: A laboratory model of molecular evolution. Proc. Natl. Acad. Sci. USA **98**:903-907.

90. Koning, S.M., Albers, S.V., Konings, W.N., Driessen, A.J.M. (2002) Sugar transport in (hyper)thermophilic archaea. Res. Microbiol. **153**:61-67.

91. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., Sonnhammer, E.L. (2002) The Pfam protein families database. Nucleic Acids Res. **30**:276-280.

92. Henikoff, S., Henikoff, J.G. (1991) Automated assembly of protein blocks for database searching. Nucleic Acids Res. **19**:6565-6572.

93. Huang, J.Y., Brutlag, D.L. (2001) The EMOTIF database. Nucleic Acids Res. **29**:202-204.

94. Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., Bork, P. (2000) SMART: a web-based tool for the study of genetically mobile domains. Nucleic Acids Res. **28**:231-234.

95. Attwood, T.K., Beck, M.E. (1994) PRINTS--a protein motif fingerprint database. Protein Eng. **7**:841-848.

96. Attwood, T.K., Beck, M.E., Bleasby, A.J., Parry-Smith, D.J. (1994) PRINTS--a database of protein motif fingerprints. Nucleic Acids Res. **22**:3590-3596.

97. Marchler-Bauer, A., Panchenko, A.R., Shoemaker, B.A., Thiessen, P.A., Geer, L.Y., Bryant, S.H. (2002) CDD: a database of conserved domain alignments with links to domain three-dimensional structure. Nucleic Acids Res. **30**:281-283.

98. Snel, B., Lehmann, G., Bork, P., Huynen, M.A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res. **28**:3442-3444.

99. Tatusov, R.L., Koonin, E.V., Lipman, D.J. (1997) A genomic perspective on protein families. Science **278**:631-637.

100. Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29**:22-28.

101. Frishman, D., Albermann, K., Hani, J., Heumann, K., Metanomski, A., Zollner, A., Mewes, H.W. (2001) Functional and structural genomics using PEDANT. Bioinformatics **17**:44-57.

102. Roth, F.P., Hughes, J.D., Estep, P.W., Church, G.M. (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol. **16**:939-945.

103. McGuire, A.M., Hughes, J.D., Church, G.M. (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. Genome Res. **10**:744-757.

104. Fumoto, M., Miyazaki, S., Sugawara, H. (2002) Genome Information Broker (GIB): data retrieval and comparative analysis system for completed microbial genomes and more. Nucleic Acids Res. **30**:66-68.

105. Bairoch, A. (2000) The ENZYME database in 2000. Nucleic Acids Res. **28**:304-305.

106. Goto, S., Okuno, Y., Hattori, M., Nishioka, T., Kanehisa, M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. Nucleic Acids Res. **30**:402-404.

107. Overbeek, R., Larsen, N., Pusch, G.D., D'Souza, M., Selkov, E., Jr., Kyrpides, N., Fonstein, M., Maltsev, N., Selkov, E. (2000) WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res. **28**:123-125.

108. Nielsen, H., Engelbrecht, J., Brunak, S., von Heijne, G. (1997) A neural network method for identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Int. J. Neural. Syst. **8**:581-599.

109. Hoffmann, K., Stoffel, W. (1993) TMbase - A database of membrane spanning proteins segments. Biol. Chem. Hoppe-Seyler **347**:166.

110. Schomburg, I., Chang, A., Schomburg, D. (2002) BRENDA, enzyme data and metabolic information. Nucleic Acids Res. **30**:47-49.

111. Thompson, J.D., Higgins, D.G., Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673-4680.

112. Nakai, K., Horton, P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem. Sci. **24**:34-36.

113. Krogh, A., Larsson, B., von Heijne, G., Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol. **305**:567-580.

114. Cserzo, M., Wallin, E., Simon, I., von Heijne, G., Elofsson, A. (1997) Prediction of transmembrane alpha-helices in prokaryotic membrane proteins: the dense alignment surface method. Protein Eng. **10**:673-676.

115. Haft, D.H., Loftus, B.J., Richardson, D.L., Yang, F., Eisen, J.A., Paulsen, I.T., White, O. (2001) TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. **29**:41-43.

116. Baxevanis, A.D. (2002) The Molecular Biology Database Collection: 2002 update. Nucleic Acids Res. **30**:1-12.

117. Nelson, K.E., Paulsen, I.T., Heidelberg, J.F., Fraser, C.M. (2000) Status of genome projects for nonpathogenic bacteria and archaea. Nat. Biotechnol. **18**:1049-1054.

118. Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Nakazawa, H., Takamiya, M., Masuda, S., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Kikuchi, H., et al. (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res. **6**:83-101, 145-152.

119. (2001) Direct Submission: *Pyrococcus abyssi* genomic sequence. National Center for Biotechnology Information, NIH, Bethesda, MD 20894.

120. Fitz-Gibbon, S.T., Ladner, H., Kim, U.J., Stetter, K.O., Simon, M.I., Miller, J.H. (2002) Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. Proc. Natl. Acad. Sci. USA **99**:984-989.

121. Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G., Lenox, A.L., Graham, D.E., Overbeek, R., Snead, M.A., Keller, M., Aujay, M., Huber, R., Feldman, R.A.,

Short, J.M., Olsen, G.J., Swanson, R.V. (1998) The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature **392**:353-358.

122. Nelson, K.E., Clayton, R.A., Gill, S.R., Gwinn, M.L., Dodson, R.J., Haft, D.H., Hickey, E.K., Peterson, L.D., Nelson, W.C., Ketchum, K.A., McDonald, L., Utterback, T.R., Malek, J.A., Linher, K.D., Garrett, M.M., Stewart, A.M., Cotton, M.D., Pratt, M.S., Phillips, C.A., Richardson, D., Heidelberg, J., Sutton, G.G., Fleischmann, R.D., Eisen, J.A., White, O., Salzberg, S.L., Smith, H.O., Venter, J.C., Fraser, C.M. (1999) Evidence for lateral gene transfer between archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature **399**:323-329.

123. Voorhorst, W.G.B., Rik, I.L., Luesink, E.J., Devos, W.M. (1995) Characterization of the celB gene coding for beta-glucosidase from the hyperthermophilic archaeon *Pyrococcus furiosus* and its expression and site-directed mutation in *Escherichia coli*. J. Bacteriol. **177**:7105-7111.

124. Kengen, S.W.M., Luesink, E.J., Stams, A.J.M., Zehnder, A.J.B. (1993) Purification and characterization of an extremely thermostable beta-glucosidase from the hyperthermophilic archaeon *Pyrococcus furiosus*. Eur. J. Biochem. **213**:305-312.

125. Verhees, C.H. (1999) Direct Submission: *Pyrococcus furiosus* genomic sequence. Laboratory of Microbiology, Wageningen University and Research Center, Hesselink van Suchtelenweg 4, Wageningen NL-6703 CT, The Netherlands.

126. Bauer, M.W., Driskill, L.E., Callen, W., Snead, M.A., Mathur, E.J., Kelly, R.M. (1999) An endoglucanase, eglA, from the hyperthermophilic archaeon *Pyrococcus furiosus* hydrolyzes beta-1,4 bonds in mixed-linkage (1 -> 3),(1 -> 4)-beta-D-glucans and cellulose. J. Bacteriol. **181**:284-290.

127. Bauer, M.W., Bylina, E.J., Swanson, R.V., Kelly, R.M. (1996) Comparison of a beta-glucosidase and a beta-mannosidase from the hyperthermophilic archaeon *Pyrococcus furiosus*. Purification, characterization, gene cloning, and sequence analysis. J. Biol. Chem. **271**:23749-23755.

128. Gueguen, Y., Voorhorst, W.G.B., van der Oost, J., deVos, W.M. (1997) Molecular and biochemical characterization of an endo-beta-1,3-glucanase of the hyperthermophilic archaeon *Pyrococcus furiosus*. J. Biol. Chem. **272**:31258-31264.

129. Tanaka, T., Fujiwara, S., Nishikori, S., Fukui, T., Takagi, M., Imanaka, T. (1999) A unique chitinase with dual active sites and triple substrate binding sites from the hyperthermophilic archaeon *Pyrococcus kodakaraensis* KOD1. Appl. Environ. Microbiol. **65**:5338-5344.

130. Jorgensen, S., Vorgias, C.E., Antranikian, G. (1997) Cloning, sequencing, characterization, and expression of an extracellular alpha-amylase from the hyperthermophilic archaeon *Pyrococcus furiosus* in *Escherichia coli* and *Bacillus subtilis*. J. Biol. Chem. **272**:16335-16342.

131. Savchenko, A., Vieille, C., Kang, S., Zeikus, J.G. (2002) *Pyrococcus furiosus* alpha-amylase is stabilized by calcium and zinc. **41**:6193-6201.

132. Laderman, K., Davis, B., Krutzsch, H., Lewis, M., Griko, Y., Privalov, P., Anfinsen, C. (1993) The purification and characterization of an extremely thermostable alpha-amylase from the hyperthermophilic archaebacterium *Pyrococcus furiosus.* J. Biol. Chem. **268**:24394-24401.

133. Dong, G.Q., Vieille, C., Zeikus, J.G. (1997) Cloning, sequencing, and expression of the gene encoding amylopullulanase from *Pyrococcus furiosus* and biochemical

characterization of the recombinant enzyme. Appl. Environ. Microbiol. **63**:3577-3584.

Figure 1. Biochemical activities of Man5 and Cel5A in *Thermotoga maritima*. Shown is the biochemical comparison of the activity of each enzyme against galactomannan, glucomannan and beta-glucan as well as the genomic position of the genes encoding the enzymes. (Based on data in ref. 42).

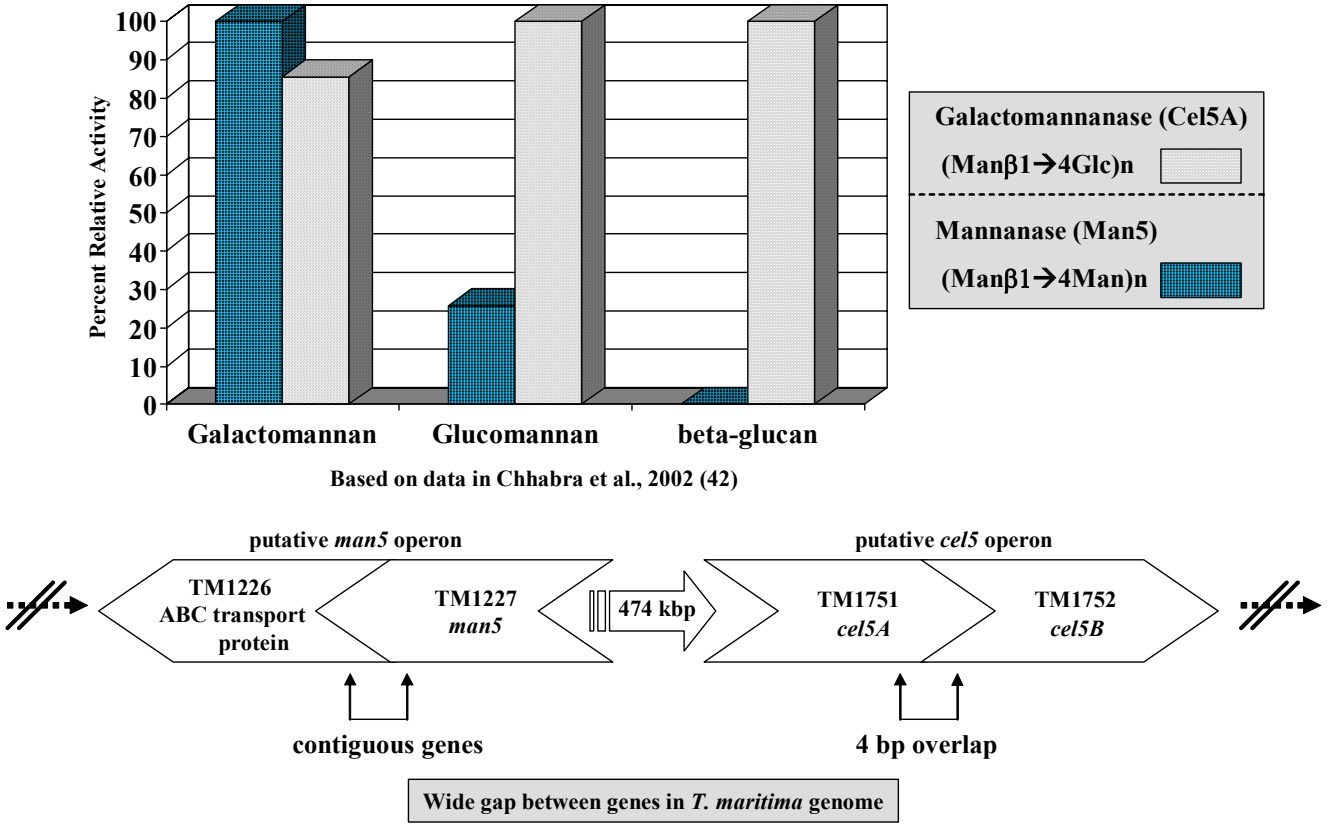


Figure 2. Putative genes encoding proteins important in sugar metabolism in *Pyrococcus furiosus* represented by all known intracellular and extracellular glycosidases and ABC transporters identified from genome sequence information and the online CAZY database (http://afmb.cnrs-mrs.fr/~cazy/CAZY/), which classifies glycosidases according to the family-based scheme of Henrisatt et al., 1998 (25). *Known and putative ABC transport proteins from the *P. furious* genome are given, including those in the annotation listed as belonging to the sugar transport- or carbohydrate uptake transporter (CUT)-family and the di/oligopeptide transport- (Opp-) family (90). Transporters that are characterized or are located close to known or putative glycosidases on the *P. furiosus* genome are shown in italics. OM, outer membrane; CM, cell membrane; PTS, phosphotransferase system; Pi, phosphate.

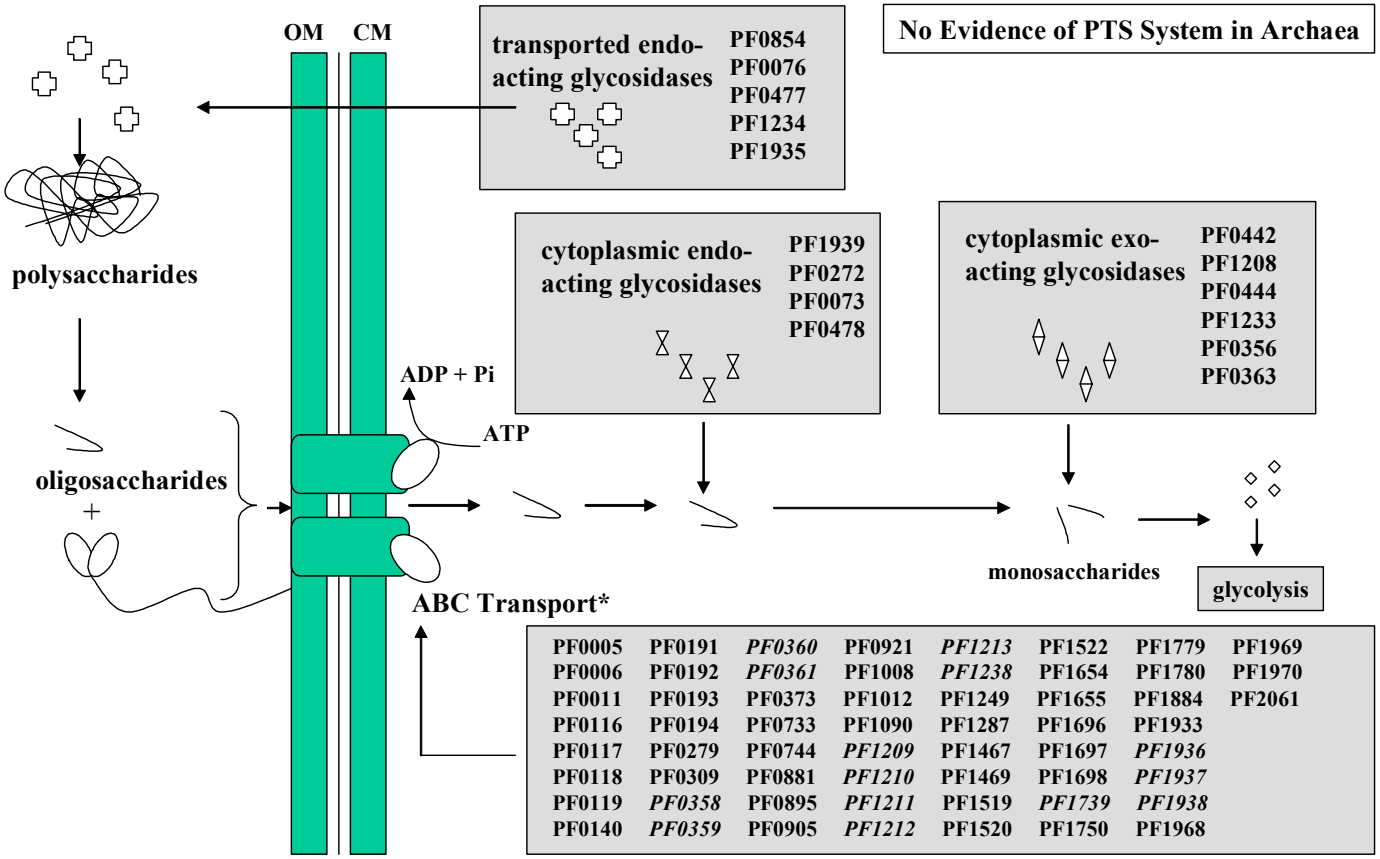


Table 1. Useful Bioinformatic Tools for Microbial Systems[1]

Search Tool	URL	Description
Protein-Sequence Databases		
BLAST	http://www.ncbi.nlm.nih.gov/BLAST/	Basic Local Alignment Search Tool; allows the rapid comparison of sequences (34); a fast way to compare sequences in a database, identify gene or protein sequences, and identify regions of similarity between a sequence of interest and those in a database; Variations on BLAST, such as PSI-BLAST or Gapped-BLAST, allow higher sensitivity.
PROSITE	http://ca.expasy.org/prosite/	Helps elucidate the function of an unknown protein sequence (translated from cDNA or genomic sequence) by comparison with known families of proteins (53).
Pfam	http://pfam.wustl.edu/hmmsearch.shtml	A database of multiple protein domain alignments capable of assessing and identifying proteins with multiple domains, detecting end-to-end similarity among protein sequences (91).
Blocks	http://blocks.fhcrc.org	Detects local regions of similarity in proteins (92).
eMOTIF	http://motif.stanford.edu/emotif	Determines and searches for protein motifs (93).
SMART	http://smart.embl-heidelberg.de	Simple Modular Architecture Research Tool allows the analysis of domain architecture (94).
PRINTS	http://www.bioinf.man.ac.uk/dbbrowswer/PRINTS/	Compendium of conserved protein motif sets (95, 96).
CDD	http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	Conserved Domain Database; composed of multiple sequence alignments for conserved regions of proteins (97).
TOPITS	http://www.embl-heidelberg.de/predictprotein/predictprotein.html	Prediction-based threading program, useful for relating sequence motifs to protein structure/function (53).
Genome Comparisons		
STRING	http://www.bork.embl-heidelberg.de/STRING/	Search tool for recurring instances of neighboring genes able to locate and display the genes that repeatedly occur in clusters on published genome sequences (98); genes that occur in repeated clusters across a wide range of genomic sequence often indicate a functional association.
COG	http://www.ncbi.nlm.nih.gov/COG/	Clusters of Orthologous Groups of proteins, determined by comparing protein sequences in complete genomes; orthologs typically have the same function (99, 100).

Table 1 (cont.)

PEDANT	http://pedant.gsf.de/	High-throughput processing of genomic data to assign functional and structural categories to proteins using a wide range of bioinformatic approaches (101).
AlignAce	http://arep.med.harvard.edu/mrnadata/mrnasoft.html	Aligns Nucleic Acid Conserved Elements; predicts functional interactions based on comparative genomics (102, 103).
Genome Information Broker	http://gib.genes.nig.ac.jp	Allows the retrieval and visualization of regions of any sequenced microbial genome that are of interest along with biological annotation (104).
Metabolic Databases		
EcoCyc	http://ecocyc.org	Annotation on all known metabolic and signal transduction pathways for *E. coli*, including a description of the genome and biochemical machinery of the organism (65).
ENZYME	http://www.expasy.org/enzyme/	Provides nomenclature, catalytic activities, and cofactors associated with an enzyme of interest (105).
LIGAND	http://www.genome.ad.jp/ligand/	Composed of three main areas; designed to provide information which links biological and chemical aspects of life; COMPOUND gives information about metabolites and associated chemical compounds; REACTION is for the collection of metabolic reactions; ENZYME gives all known enzymatic reactions pertaining to the protein of interest (106).
KEGG	http://www.genome.ad.jp/kegg	Kyoto Encyclopedia of Genes and Genomes; provides functional information derived from genome sequence data (64).
WIT2	http://wit.mcs.anl.gov/WIT2/	What is There database; contains information on metabolic pathways and is based on sequence comparisons, biochemical, and phenotypic data (107).
Other Useful Tools		
SignalP	http://www.cbs.dtu.dk/services/SignalP/	Identification of signal peptides and cleavage sites based on neural networks for prokaryotic and eukaryotic systems (108).
TMPRED	http://www.ch.embnet.org/software/TMPRED_form.html	Predicts membrane-spanning regions of proteins and orientation; based on the statistical algorithms of TMbase (109).
BRENDA	http://www.brenda.uni-koeln.de/	Collection of enzyme functional data (110).

Table 1 (cont.)

CLUSTALW	http://www.ebi.ac.uk/clustalw/	Multiple sequence alignment program for DNA or proteins that allows sequences to be aligned in order to view similarities or differences between molecules (111).
PSORT	http://psort.nibb.ac.jp/form.html	Prediction of protein sorting signals, or protein localization sites in cells, from amino acid sequence data (112).
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/	Can predict the occurrence of transmembrane helices; based on Hidden Markov Models (113).
DAS	http://www.sbc.su.se/~miklos/DAS/	Dense Alignment Surface method; predicts transmembrane regions from amino acid sequence data for any integral membrane protein (114).
TIGRFAMs	http://www.tigr.org/TIGRFAMs/	Functional identification of proteins; based on Hidden Markov Models (115).

[1]See Baxevanis, 2002 (116) for a more extensive listing of bioinformatic databases. See also Nelson et al., 2000 (117).

Table 2. Available Genome Sequences for Hyperthermophilic Microorganisms

Name	Year	Description	T_{opt} (°C)	Genome Size (Mbp)	ORFs	Genes w/unknown Function[1]	Unique Genes[1]	% G+C	Isolation site	Ref.
Archaea										
Aeropyrum pernix	1999	Strictly aerobic crenarchaeon	95	1.67	2,694	523 (19%)	1,538 (57%)	56	Kodakara	(118)
Archaeoglobus fulgidus	1997	Strictly anaerobic Archaeoglobales, sulfur-metabolizing	83	2.18	2,436	1,315 (54%)	641 (25%)	49	Vulcano	(44)
Methanococcus jannaschii	1996	Anaerobic, autotrophic, Methanogenic Methanococcales	85	1.66	1,729	1,076 (62%)	525 (30%)	31	East Pacific Rise	(45)
Pyrococcus abyssi	2001	Anaerobic Thermococcales	98	1.77	1,765	---	---	45	North Fiji Basin	(119)
Pyrococcus horikoshii	1998	Anaerobic, obligately heterotrophic Thermococcales	98	1.74	2,061	859 (42%)	453 (22%)	42	Okinawa Trough	(17)
Pyrococcus furiosus	2002	Anaerobic Thermococcales, grows well on sugars and peptides	98	1.91	2,208	---	---	40	Vulcano	(43)
Sulfolobus solfataricus	2001	Aerobic Solfolobales, grows best at low pH	80	2.99	3,032	577 (22%)	743 (25%)	---	Pisciarelli Solfatara	(23)
Pyrobaculum aerophilum	2002	Facultatively aerobic nitrate-reducing crenarchaeon	100	2.22	2,587	---	302 (12%)	51	Maronti Beach	(120)
Bacteria										
Aquifex aeolicus	1998	Microaerophilic Aquificaceae, Obligate chemolithoautotroph	95	1.55	1,512	663 (43%)	407 (27%)	43	Not reported	(121)
Thermotoga maritima	1999	Anaerobic Thermotogales, metabolizes simple and complex carbohydrates	80	1.86	1,877	863 (43%)	373 (20%)	46	Vulcano	(122)

[1]Refers to the time of genome sequence publication

Table 3. Proteases in *P. furiosus*

	Locus	S[1]	Nuc.	a.a.	Ph	Pa	Mj	Af	Pae	Ap	Ss	Tm	Aa
ATP-Dependent Proteases													
Proteasome, subunit beta (PsmB-1)	PF1404	N	621	206	x	x	x	x	x	x	x		
Proteasome, subunit beta (PsmB-2)	PF0159	N	599	199	x	x	x	x	x	x	x		
ATP-dependent Regulatory Subunit (PAN)	PF0115	N	1199	399	x	x	x	x		x	x		
ATP-dependent LA (Lon)	PF0467	N	3140	1046	x	x		x					
Proteasome, subunit alpha (PsmA)	PF1571	N	798	265	x	x	x	x	x	x	x		
ATP-Independent Proteases													
Subtilisin-like protease	PF0688	N	593	197					x	x[2]			
Intracellular protease I (PfpI)	PF1719	N	582	193	x	x	x	x	x	x	x		x
Periplasmic serine protease, putative	PF0240	Y	842	280	x	x	x	x		x		x	x
Metalloprotease	PF0392	N	1253	417	x	x		x		x			
Alkaline serine protease	PF1670	Y	1992	663				x	x	x			
Metalloprotease	PF0167	Y	1133	377	x	x	x	x		x			
Putative bacteriocin/protease	PF1191	N	785	261									
Pyrolysin	PF0287	Y	4238	1412					x[2]				
Hydrogenase maturation protease (hyc I)	PF0617	N	485	161	x	x	x	x					
Hypothetical protein	PF0760	N	1022	340	x						x	x	
Protease IV	PF1583	Y	990	329	x	x	x						x
Putative Protease	PF1905	Y	1332	443				x[2]					
Metalloprotease	PF0457	N	629	209		x	x		x		x		x
Peptidases													
Acetylornithine deacetylase (ArgE)/peptidase	PF1185	N	1061	353	x	x		x					
HtpX heat shock protein	PF1135	N	875	291	x	x	x	x		x			x

Table 3 (cont.)

Proline dipeptidase-related protein	PF0702	N	521	173	x	x						x	
protein similar to endo-1,4-β-glucanase (ytoP)	PF1861	N	1040	346	x	x	x	x	x	x		x	
D-aminopeptidase	PF1924	N	1098	365	x	x							
Signal sequence peptidase I, SEC11	PF0313	N	264	87	x	x			x	x			
Hypothetical protein	PF0669	N	932	310	x	x							x[2]
O-sialoglycoprotein endopeptidase (gcp-2)	PF0473	N	680	226	x	x		x	x	x	x		
O-sialoglycoprotein endopeptidase (gcp-1)	PF0172	N	974	324	x	x	x	x	x	x	x	x	x
Succinyl-diaminopimelate desuccinylase/peptidase	PF2048	N	1343	447	x	x				x			x[2]
Pyroglutamyl-peptidase I	PF1299	N	654	217	x	x					x		
XAA-Pro dipeptidase (proline dipeptidase)	PF1343	N	1047	348	x	x	x	x	x	x	x	x	x
Protein similar to acylaminoacyl peptidase (acylamino acid-releasing enzyme homolog)	PF0318	N	1862	620	x	x			x	x			
Prolyl endopeptidase	PF0825	N	1863	620	x	x							
Endoglucanase (CelM)/aminopeptidase	PF1547	N	1046	348	x	x	x	x	x	x	x[2]	x	
Methionine aminopeptidase (MAP) (Pep M)	PF0541	N	887	295	x	x	x	x	x	x	x	x	x
Putative proline dipeptidase	PF0747	N	1076	358	x	x	x						
heat shock protein X	PF1597	Y	800	266	x	x							
Carboxypeptidase I	PF0456	N	1499	499	x	x			x	x	x		
endoglucanase/peptidase	PF0369	N	998	332	x	x	x	x	x	x		x	
Putative aminopeptidase	PF2059	Y	1704	567									
Putative aminopeptidase	PF2063	Y	1755	584									
Putative aminopeptidase	PF2065	Y	1776	591									
Membrane dipeptidase	PF0874	N	1140	379	x	x				x			

Proteins in database were considered to be present if they contained >30% identity over >50% length of protein in database.

[1]S – indicates the presence of a signal peptide as identified from Signal P (108).

[2]predicted amino acid length of protein encoded by ORF differs significantly from predicted *P. furiosus* protein length.

P. furiosus (strain DSM3638) was compared against: Ph – *Pyrococcus horikoshii* OT3, Pa – *Pyrococcus abyssi* GE5, Mj - *Methanococcus jannaschii* DSM 2661, Af – *Archaeoglobus fulgidus* VC-16, Pae – *Pyrobaculum aerophilum* IM2, Ap – *Aeropyrum pernix* K1, Ss – *Solfolobus solfataricus* P2, Tm – *Thermotoga maritima* MSB8, Aa – *Aquifex aeolicus* VF5.

Table 4: Glycosidase Inventory from *P. furiosus* based on Genomic Sequence Data

Locus	Annotation	Reference	Activity	S[1]	Ph	Pa	Mj	Af	Pae	Ap	Ss	Tm	Aa
	Cellulases												
PF0073	β-glucosidase	(123, 124)	Cel1A		x								
PF0442	β-glucosidase	(125)	Cel1B						x				
PF0854	endo-1,4-β-glucanase	(126)	Cel12	Yes								**x**	
	Mannosidases												
PF1208	β-mannosidase	(127)	Man1		x	x							
	Laminarinases												
PF0076	endo-β-1,3-glucanase	(123, 128)	Lam16	Yes								**x**	
	Chitinases												
PF1234	putative chitinase	(129)	Chi18A	Yes									
PF1233	putative chitinase	(129)	Chi18B										
	Amylases/Pullulanases												
PF0477	α-amylase	(130, 131)	Amy13	Yes			**x**					x	
PF0272	α-amylase	(132)	Amy57A		x	x	**x**		x				x
PF1935	amylopullulanase	(133)	Amy57B			x			x		x		
PF0478	α-amylase[2]		Amy13								x		x
PF1939	neopullulanase		Pul13						x		x	x	

Table 4 (cont.)

	Galactosidases					
PF0444	α-galactosidase	Gal57	x			
PF0356	β-galactosidase[2]	Gal1		x		**x**
PF0363	β-galactosidase precursor[2]	Gal35	x	x		

[1]S – indicates the presence of a signal peptide as identified from Signal P (108); [2]putative proteins.
Characterized proteins are noted in bold.
P. furiosus (strain DSM3638) was compared against: Ph – *Pyrococcus horikoshii* OT3, Pa – *Pyrococcus abyssi* GE5, Mj - *Methanococcus jannaschii* DSM 2661, Af – *Archaeoglobus fulgidus* VC-16, Pae – *Pyrobaculum aerophilum* IM2, Ap – *Aeropyrum pernix* K1, Ss – *Sulfolobus solfataricus* P2, Tm – *Thermotoga maritima* MSB8, Aa – *Aquifex aeolicus* VF5.

Chapter 3: Estimating genome-wide transcriptional variation within and between steady states for continuous growth of the hyperthermophile *Thermotoga maritima*

Keith R. Shockley[1], *Kevin L. Scott*[2], *Marybeth A. Pysz*[3], *Shannon B. Conners*[1], *Matthew R. Johnson*[1], *Clemente I. Montero*[1], *Russ Wolfinger*[2] *and Robert M. Kelly*[1]*

[1] Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

[2] SAS Institute Inc.
100 SAS Campus Drive
Cary, NC 27513

[3] Current address: Roswell Park Cancer Institute
Department of Pharmacology and Therapeutic
Elm and Carlton Streets
Buffalo, NY 14263

To be Submitted to: ***Nucleic Acids Research*** (May, 2004)

Running Title: Transcriptional variation in *Thermotoga maritima*

*Address inquiries to: **Robert M. Kelly**
Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

Phone: (919) 515-6396
Fax: (919) 515-3465
Email: rmkelly@eos.ncsu.edu

Abstract

Whole genome spotted cDNA microarrays were used to evaluate transcriptional variation within and between steady states during continuous growth of the hyperthermophilic bacterium *Thermotoga maritima*. An experimental loop design facilitated time-dependent assessment of gene expression variability as a function of growth rate (0.17 h^{-1}, 0.25 h^{-1}) and temperature (80°C, 85°C). Mixed models, in which effects were partitioned into sources of technical and biological variation, were used to assess statistical significance for estimating differential gene expression. In total, 320 transcripts responded to growth rate changes while 102 transcripts responded to temperature perturbations at a Bonferroni adjusted significance level of 0.05. A small but consequential number of ORFs were found to have variances sufficiently large to result in statistically significant, but experimentally irrelevant, inflections in gene expression within a particular biological context (fixed growth rate and temperature). Of the 422 ORFs that were differentially regulated between steady states, 93 ORFs also showed significant variability in gene expression within a steady state. Furthermore, a significant number of ORFs exhibited gene expression effects arising from dye type (4.1%), sampling time (3.7%), or treatment-time interactions (1.1%). The results reported here demonstrate the utility of mixed effects models for partitioning random and systematic errors, treatment effects, sampling strategy as well as gene expression variability in cDNA-based transcriptional response analysis.

Introduction

Establishing the connection between genome sequence data and biological function remains the central challenge of the post-genomics era. To address this challenge, transcriptional response analysis using cDNA microarrays has been shown to provide unprecedented insights into the complex patterns of gene expression that follow the exposure of an organism or cell to environmentally different conditions (1, 4, 28). In some cases, differential expression of specific genes may seem readily apparent because of the relative magnitude of response. In other instances, subtle changes in gene expression patterns underlying biologically important phenomena, may be difficult to discern and thus be ignored. In the final analysis, no matter what the magnitude of differential expression happens to be, appropriate statistical measures are needed to properly assess the significance of response information. This necessitates that statistical variances arising from differential response experiments be partitioned in such a way that biologically significant responses are not confounded with systematic and random experimental error nor diluted by other irrelevant factors.

A potential source of error frequently ignored in global transcriptional analysis is the fluctuation in gene expression that arises within a specific environmental context. While gene expression variability studies for individual genes have been reported (19), this issue is typically not considered for large sets of genes, as is the case with entire genome transcriptional analysis. It would be best to eliminate this source of biological variability when comparisons in gene expression levels for small but statistically significant changes are made (21). This should be done so that insights into biological mechanisms can be discerned at the most meaningful level of resolution. Previous work has aimed at minimizing biological

variability in gene expression within a given context by sample pooling or by multiple comparisons of samples taken from independent batch or continuous culture experiments (9, 21). However, this may mask small but significant biological variation in transcription, which can depend strongly on the size or source of the pools being sampled. Thus, much of the variation between experimental conditions could reflect the variation within an experimental condition.

There have been many statistical methodologies proposed to interpret biologically significant phenomena from cDNA microarray-based transcriptional response experiments. For example, simple fold-changes based on differences in fluorescence intensities (e.g., the "2-fold change" rule) or their associated test statistics (e.g., "p-values") have been used to ascertain differential expression (28, 29, 32). However, more statistically robust methodologies are often necessary to fully appreciate the basis for changes in the transcriptome. In this regard, some recent studies have adopted mixed effects models for differential gene expression analysis that partition different sources of variance for individual genes (4, 10, 34). These models could be even more useful if candidates for differential gene expression are modeled by incorporating biological controls (9, 21, 33) that account for significant alterations in gene expression reflecting natural fluctuations in transcription within a given environmental context. In order to detect small but biologically meaningful changes in expression patterns, error modeling should be performed on each open reading frame under each tested growth condition. Mixed effects models can be employed to treat this aspect of microarray data analysis although this aspect has not been reported.

Additional complications in interpreting microarray data arise from the experimental methodologies used to generate samples for analysis. Most microarray experiments to date

have focused on batch growth, where environmental changes over time can be significant and phase-dependent. Alternatively, continuous culture can be used to focus on one growth condition of interest while holding other growth parameters constant. A significant drawback to physiological studies using chemostats is the potential for contamination. However, this problem is largely averted in the case of hyperthermophilic microorganisms (organisms with optimal growth temperatures of 80°C and above) (3, 22, 25). From this perspective, hyperthermophiles are ideal systems to study long term, steady-state effects on cell transcription for various physiological states. *Thermotoga maritima*, an obligately anaerobic, hyperthermophilic, heterotrophic bacterium growing optimally at 80° (8), whose genome has been fully sequenced (15), has been grown successfully in chemostat culture (26). Targeted cDNA microarrays have been used previously to investigate carbohydrate utilization patterns (4) and heat shock response (23) in this organism. Here, we use a mixed effects ANOVA model (4, 10), in conjunction with chemostat-based transcriptional response experiments with whole genome spotted cDNA microarrays, to investigate sources of variance that contribute to the observed patterns of differential gene expression both within and between mechanical steady states (temperature, dilution rate). Among the issues examined are the influence of technical errors, sampling strategy and variability within a given growth condition on the interpretation of biologically important phenomena.

Materials and methods

Microorganism and growth conditions

Thermotoga maritima (DSM 3109) was grown anaerobically on Sea Salts Medium (SSM) containing 40 g/l sea salts (Sigma, St. Louis, MO), 1 g/l yeast extract (Fisher

Scientific, Pittsburgh, PA), 3.1 g/l PIPES buffer (Sigma Chemical, St. Louis, MO), 2 g/l tryptone, 2 ml/l of 0.05% Resazurin, and 10 ml/l 10X Wolin minerals (6). Media was adjusted to pH 6.8 and autoclaved prior to use. Batch cultures (50 ml) were inoculated under N_2 (high purity nitrogen; National Welders, Raleigh, NC) headspace, as previously described (22), and were grown at 80°C for 8 to 10 hours in oil baths. Maltose (Sigma Chemical, St. Louis) was added to SSM (final concentration 5 g/l) as a carbon source prior to inoculation. Continuous cultivation of *T. maritima* was performed in a 2-L five-neck, round-bottom flask, as previously described (22). A 50 ml batch culture was used to inoculate 1 L of SSM medium supplemented with 5 g/l maltose in the flask. This seed culture was grown at 80°C for 8.5 hours under continuous nitrogen sparging, after which medium was fed at a dilution rate of 0.25 h^{-1}. Media for continuous cultivation was prepared in 9 L batches at 1.2X concentration as mentioned above, to which 1 L of a filter-sterilized maltose solution (50 g) was added immediately after autoclaving. The pH of the culture was continuously controlled with a Chemcadet pH controller (Cole Parmer, Vernon Hills, IL) by the addition of 1 M NaOH. Temperature control was performed with a Digi-Sense controller (Cole-Parmer, Vernon Hills, IL) (variations were typically ± 0.8°C), and verified by a mercury glass thermometer inserted into the culture. Steady-state conditions were monitored by following cell counts (see below) and optical densities at 600 nm. Samples for RNA extractions were collected on ice.

After approximately a week of operation, the temperature of the continuous culture was shifted to 85°C to monitor growth at supraoptimal temperatures. Cell densities were measured every hour following the temperature stabilization at 85°C to determine the new steady-state, which occurred after approximately 4 reactor volume changes.

RNA sample collection

Approximately 200 ml samples of planktonic cells were collected on ice for RNA extraction, and centrifuged at 10,000×g for 20 minutes. Pellets were rinsed twice with 300 mM NaCl and spun at 13,000 rpm for 5 minutes. RNA was isolated by an acid phenol extraction, ethanol precipitation, and purified using Promega total RNA kits (Promega, Madison WI). Concentrations and degree of purity were determined by optical density at 260 nm and 280 nm, as well as by gel electrophoresis. Total RNA was extracted from 6 different time points during steady-state operation at baseline conditions (80^{o}C, dilution rate = 0.25 hr^{-1}), 4 different time points during elevated temperature (85^{o}C, dilution rate = 0.25 hr^{-1}), and 3 different time points during slower growth (80^{o}C, dilution rate = 0.17 hr^{-1}).

Construction of the full genome DNA microarray

Open reading frames of all genes were extracted from the *T. maritima* MSB8 genome available at http://www.tigr.org/tigrscripts/CMR2/GenomePage3.spl?database=btm. DNA primers were designed with similar annealing temperatures and minimal hairpin formation using Genomax 2.0 (Informax, Bethesda, MD). Probes were PCR-amplified in a PTC-100 Thermocycler (MJ Research, Inc., Waltham, MA) using *Taq* polymerase (Boehringer, Indianapolis, IN) and *T. maritima* genomic DNA, isolated as described previously (4). After amplification, the PCR products were quantified and then purified to a concentration of 100 ng/µl using QIAquick PCR purification kits (Qiagen, Valencia, CA). Purified PCR products were re-suspended in 50% DMSO, randomized, dispensed evenly into microarray printing plates (Genetix) and printed onto GAPS II aminosilane-coated microscope slides (Corning, Corning, NY) with a QArray-Mini Arrayer (Genetix, London, UK). The DNA was attached

to the substrate by UV crosslinking in a GS GeneLinker UV Chamber (BioRad, Hercules, CA) at 250 mJ and then baked at 75^{o}C for 2 hours. The crosslinked slides were protected by storage in a desiccator and kept away from light at room temperature until they were used.

Preparation of cDNA and hybridization

First-strand cDNA was prepared from *T. maritima* total RNA using Stratascript (Stratagene) and random hexamer primers (Invitrogen Life Technologies, Carlsbad, CA) by the incorporation of 5-[3-Aminoallyl]-2'-deoxyuridine-5'-triphosphate (aa-dUTP) (Sigma) (7). The generated cDNA products were purified using the QIAquick PCR purification kit (Qiagen) and reacted with monoreactive Cyanine-3 (Cy-3) and Cyanine-5 (Cy-5) NHS-esters (Amersham Biosciences, Inc., Piscataway, NJ). Another round of purification was used to remove unincorporated dyes. The hybridizations and washes were performed as described previously (29). The slides were scanned using the Scanarray 4000 scanner (Perkin Elmer, Fremont, CA). Signal intensity data for all experiments was extracted using Scanarray (Perkin Elmer).

Calculation of simple fold change and log_2 significance

The raw intensity data from ScanArray was exported into Excel (MicroSoft, Seattle). Local background was subtracted from each spot. Any ORFs containing negative values after background subtraction (resulting from extremely low transcript abundance) were discarded from further analysis. The Cy-3 signal from the first sample taken at 103.7 hours (80^{o}C, D=0.25 hr^{-1}) was used as a basis for normalizing to the total fluorescence intensities from all other treatments and channels. Simple fold change was calculated by dividing the mean

signal intensities from one treatment by the mean signal intensities from a different treatment. To calculate $\log_2$ estimates, normalized signals were converted to a $\log_2$ scale and averaged by treatment. Differences in average $\log_2$ values for each ORF between treatments and sampling time were calculated and the results were used to calculate significance values. Bonferroni corrections were applied in order to calculate adjusted P values corresponding to $\alpha<0.05$.

Mixed model analyses

Replication of treatments, arrays, dyes, and cDNA spots allowed the use of analysis of variance (ANOVA) models for data analysis (34). Loop designs were constructed as indicated below. The data import code reported previously (4) was used to analyze spot intensities obtained from Scanarray. A linear normalization ANOVA model (34) was used to estimate global variation in the form of fixed (dye (D), treatment (T)) and random (array (A) and spot within array A(S)) effects and random error using the model $\log_2(y_{ijklm}) = A_i + D_j + (Tt)_{kl} + A_i(S)_m + \varepsilon_{ijklmn}$. Here, random effects are distributed normally with a mean of zero and variances σ^2_A and σ^2_{AS}. A gene-specific ANOVA model was then used to partition the remaining variation into gene-specific effects using the model $r_{ijklm} = \mu_n + D_{jn} + (Tt)_{kln} + A_{in} + A(S)_{imn} + \gamma_{ijklmn}$. A Bonferroni correction (adj. $P<0.05$) was used to adjust for the expected increase in false positives due to multiple comparisons (34).

Quantitative PCR

Quantitative PCR (QPCR) was used to confirm the microarray results of selected genes using pooled RNA samples. Primers were designed using Genomax software, using

Oligoanalyzer 3.0 http://biotools.idtdna.com/Analyzer/ and mFold http://biotools.idtdna.com/mFold/: Reverse transcription of RNA to cDNA was performed as described above. QPCR of cDNA was performed using the SYBRGREEN kit and iCycler iQ Real-time PCR detection system (Bio-Rad Laboratories, Hercules, CA) according to manufacturer protocols. Briefly, reactions for 10 ng of samples were carried out for the 5 genes at three different temperatures to determine the optimum S-curves. Optimization indicated that all reactions could be performed at 55°C. Standard curves (20 ng, 4 ng, 0.8 ng, and 0.16 ng) for each gene. Quantitative results were calculated using vendor-provided software (Bio-Rad Laboratories, Hercules, CA). QPCR results for two ORFs (TM0403 and TM1591) were consistent with those obtained from the mixed model analysis. QPCR indicated that transcript abundance for TM0403 (encoding a nitrogen regulatory protein), one of the most strongly regulated genes in the entire pooled RNA experiment (adjusted P-value = 3.4×10^{-9}), was +4.3-fold higher for growth at 80°C, D=0.17 hr^{-1} than at 80°C, D=0.25 hr^{-1}. For the same treatment comparison, TM1591 (encoding ribosomal protein L35), (adjusted P-value = 1), yielded a QPCR fold change of +1.1.

For complete information on significance of expression changes, fold changes, pair-wise volcano plots, and hierarchical clustering for all of the genes included on the array, see our website, http:///www.che.ncsu.edu/extremophiles/ (which will provide these data upon acceptance of the manuscript).

Results and discussion

The objective of this study was to evaluate sources of variance that can arise from cDNA microarray-based investigations of differential gene expression patterns and assess their impact on the interpretation of biological response. Comparisons were made between two statistical methodologies that can be used for microarray analysis: simple fold changes (directly derived from normalized fluorescence intensities) were contrasted with probability estimates resulting from difference estimates of both $\log_2$ transformed data and mixed effects ANOVA models. Side-by-side comparisons of these approaches were used to determine the impact that systematic and random errors have on information obtained from transcriptional response experiments.

Also of interest was capturing the influence that biological variation, present within a given environmental context, had on interpretation of biological phenomena. Most gene expression studies involving microbial cultures are conducted in batch mode in which pre-cultures are first grown overnight, then inoculated into replicate flasks and sampled after growth has reached a certain cell density. However, these cultures are not metabolically, physiologically or dynamically identical. Duplicate cultures should exhibit identical growth dynamics but, even under the most carefully reproduced circumstances, different cultures may not reach the same density in the same amount of time. This may be due to inconsistent experimental conditions (e.g. temperatures or mixing of microenvironments), slight modulations in inoculation procedures, or stochastic processes within individual cultures. In batch cultures the environment is constantly changing as nutrients are being depleted from the medium. Continuous cultures permit the physiological examination of microorganisms by allowing the investigation of one variable at a time while maintaining an otherwise

unchanging environment with a constant growth rate (2, 27). The continuous culture technique works particularly well for hyperthermophilic organisms due to the small contamination risk (3, 25). In continuous cultures, cell growth rate can be modulated through changing the dilution rate. Here, using increased replication, mixed model ANOVA analyses were used to differentiate systematic and random errors from biologically significant effects. This approach shows the utility of microarray technology to detect small differences in transcription that were obscured with methods that did not take into account important experimental effects. To better assess biological variation within a continuous culture, independent hybridizations of samples collected at different time points within and between three different mechanical steady states were performed.

Experimental design

A loop experimental design (Figure 1) was used in order to examine differential gene expression both within and between steady states during continuous cultivation of *T. maritima* grown on maltose as the primary carbon and energy source. This design allowed efficient comparison of different RNA samples without the drawbacks of using a single reference sample (10). The level of replication and subsampling within array was sufficient for the testing of statistical hypotheses of treatment effects for individual genes, so that statistical tests did not rely on technical variance alone. The error degrees of freedom (119 in most cases) comparing treatment effects of individual genes were also adequate for estimating standard errors. Also, the conservative Bonferroni multiple test correction was applied to the P-values of the statistical tests to protect against false positive inferences of treatment effect. The whole genome cDNA microarray used in this study contained 1907

ORFs, or more than 99% of the genes presently predicted to be present in the *T. maritima* genome. Three distinct mechanical steady states were investigated by interrogating gene expression patterns from multiple time course samples within each growth condition: 80^{o}C with a dilution rate, D = 0.25 hr^{-1} (6 samples); 85^{o}C, D = 0.25 hr^{-1} (4 samples); 80^{o}C, D = 0.17 hr^{-1} (3 samples) (see Figure 2). These conditions were chosen so that distinct transcriptional patterns could be observed for normal growth conditions for this organism.

Normalization procedures

Absolute signal intensities were normalized in order to account for overall differences in population fluorescence between hybridized samples so that meaningful biological inferences could be drawn (24). The detection limits of microarray scanners can lead to scatter at low ends of signal intensity data, thus producing unreliable results (16). However, discarding data for signal intensities that are below an arbitrary threshold can unnecessarily eliminate valuable microarray data for which variability in fluorescence intensities is either low relative to the mean or high in one channel but relatively low in another channel. Therefore, low signal intensity data were retained here throughout all normalization procedures. Experimental error was minimized by using six replicate measurements per treatment for each gene. Different sources of variance were compared through mixed model analysis; global and gene-specific calculations took into account array and spot effects within and across genes. Effects due to array (residual = 0.3759) and spot within array (residual = 0.1008) were small compared to residuals from the global mixed model (2.7621) used for normalization, indicating that the technical arraying procedures are highly reproducible. Pin effects were negligible (data not shown) and were excluded from the final analyses.

The gene-specific normalization model outputs of interest are the least squares mean estimates, which are the quantities used for subsequent statistical tests. Probabilities associated with differences between estimates provide statistical support for assessing significance of gene expression differences between comparisons of interest. In order to ensure that probabilities do not yield false positives, a Bonferroni correction was applied to the P-value calculated from gene specific estimate differences. Because a total of 148,368 total comparisons were made in the full loop design (Figure 1), the Bonferroni significance level corresponding to $\alpha_{BON}=0.05$ was $\alpha=3.37 \times 10^{-7}$. Dye, array, and spot effects are confounded with treatment effects in the $\log_2$ transformed estimates (or the unnormalized case), but mixed model analyses were used to parse these important sources of error from treatment effects.

Figure 3A shows least squares mean (LSM) estimates resulting from the gene-specific $\log_2$ transformed and mixed model analyses; each data point represents the expression due to a unique gene-specific treatment effect. To illustrate with an example, the LSM estimate for TM1183, a putative oxidoreductase, for the 5th replicate sample from the steady state corresponding to 80^{o}C, D=0.25 hr^{-1} is 0.0 (corresponding to an average expression across all genes, average LSM = 0 ± 1.5) for the mixed model analysis and 15.5 (considerably above the average expression of 12.3 ± 1.5) for the $\log_2$ transformed analysis. The simple $\log_2$ transformation analysis resulted in TM1183 being designated as down-regulated (adj. P<0.05) in response to increased temperature and decreased growth rate and up-regulated compared to other samples within the sample steady state (growth at 80^{o}C, D=0.25 hr^{-1}). In contrast, the same ORF was only found to be up-regulated in response to decreased growth rate after mixed model analysis.

As shown in the highlighted regions in Figure 3A, the magnitude of signal intensity (indicated by percentile rankings of average normalized fluorescence intensities) is generally reflected in the resulting LSM estimates. That is, low signal intensity data usually lead to reduced LSM values and high signal intensities most often result in high LSM values, both for $\log_2$ transformed data and mixed model analyses. Figure 3A also reveals that considerably more scatter exists in the correlation between $\log_2$ transformed and mixed model LSM values resulting from low signal intensities than for high fluorescence. Low fluorescence intensities have led to increased variation in ratio data (16) or arbitrary signal intensities (28, 29) on numerous occasions in the literature. Even so, standard errors associated with least square means were relatively high across all gene ranks and LSM estimates retained uniform standard errors in most instances (Figure 3B). As shown in Figure 3B, there were a significant number of differentially expressed ORFs with high levels of standard error for ORFs with relatively high signal intensities as well as for those with low signal intensities. This is significant, because genes showing little experimental variation usually have the most significant differences in expression between treatments and those with the largest experimental variation most often yield significant treatment effects, as has been shown previously (18). Therefore, with sufficient replication and adequate statistical assessment, low signal intensity information may have higher variation that still produces meaningful results. This lends further support for retaining low signal intensity data.

Simple "fold change" criteria and estimate probabilities

Inferences of differential gene expression between different cell populations are usually based on simple fold change criteria, oftentimes extrapolated backward from data

transformed estimates (e.g., the "two-fold change rule"). In this study, however, instead of extrapolating from $\log_2$ estimates, simple fold changes were calculated directly from signal intensities between comparisons of interest after a weighted normalization based on total signal intensity. This was done in order to compare estimate differences to the variations in transcript abundance inferred directly from fluorescence intensity ratios. These simple fold changes were then compared with Bonferroni adjusted P values (adj. $P<0.05$) resulting from differences of $\log_2$ estimates derived from either $\log_2$ transformed data or the mixed model analyses. As shown in Table 1, similar numbers of significant genes were detected for fold change ranges between both estimate-based models, corresponding to fold change intervals ranging from less than 1.25 to well over 100. Estimates resulting from $\log_2$ transformed data were implicitly larger for larger fold changes (see Figure 4A) while estimates resulting from mixed model analyses did not show the same fold change dependence on magnitude (Figure 4B). When LSM values were calculated directly from $\log_2$ transformations of the data, significances at the Bonferroni corrected level that were derived from the associated estimate differences were strongly dependent on the magnitude of the fold change (see Figure 4A). However, after mixed model normalization, significant ORFs were not found to depend on the magnitude of the fold change (Figure 4B). Thus, the determination of whether a measured change in expression reflects a true biological alteration depends on the amount of variation present within a system and not strictly on fold change level. In fact, only one ORF with a fold change measurement over 100-fold (out of 12 different genes expressed over 100-fold) was significant in the mixed model analyses after a Bonferroni adjustment was applied ($P_{adj} < 0.05$); nearly the same percentage (12.1%) of the fold changes less than 1.25-fold were found to be significant at this level. This trend contrasts with the $\log_2$ transformed significance

results (not normalized for dye, array, and spot effects), which shows a very high agreement with the raw fold change data for fold changes greater than 10 (Figure 5). This indicates that simple fold change, by itself, is not a sound basis for determining differential gene expression. Furthermore, while there was 25.7% overall agreement between adj. P values (< 0.05) from the $\log_2$ transformed data and the results from simple fold changes, there was only 17.0% agreement between the mixed model and simple fold change. However, there was 23.9% agreement between differences in estimates between the $\log_2$ transformed and mixed model normalization.

Effects of sample pooling

Sometimes, due to difficulty in generating enough material, RNA samples need to be pooled in order to be able to perform the experimental analysis. To circumvent this problem, or in an attempt to reduce biological variability in samples, RNA samples are pooled together before conducting a microarray experiment. Pooling RNA samples to reduce systematic fluctuations in transcription (i.e. the smoothing effect (11, 12)), will mask small but significant biological variation. From a statistical standpoint, it has been assumed that RNA pooling may be acceptable as long as the transcriptional variation among pooled samples is considerably lower than the variation present among the biological conditions of interest (12). Even though an outlying RNA sample may bias the entire sample pool, levels of biological and technical variation are usually not known in advance and the cost of performing "pilot" studies for estimating such variability may not be practical.

From statistical considerations alone, pooling RNA samples should, by definition, affect data analysis and inference (11, 12, 20). Recent studies have considered the statistical

implications of sample pooling using virtual pooling strategies (20) or the performance of different estimators for designs with and without pooling (11). Both of these theoretical analyses predict that pooling RNA samples may be a cost-effective and useful way to allow the estimation of gene expression differences in microarray experiments, but the effects of RNA pooling have never been rigorously tested using experimental data.

The effect of sample pooling was considered here using a 3-sample/3-treatment loop design (Figure 6.A). When RNA samples are pooled, the biological variability of transcripts within steady states is necessarily confounded with the pool and inferences can only be drawn between pools (or between steady states). However, when considered separately, both the biological and technical variability within a steady state can be estimated. There was 87% agreement between pooled and unpooled samples (numbers of significant genes in the pooled sample that were significant in at least one unpooled comparison) when 3 RNA samples from each steady state were either pooled together or considered separately (see Figure 6.B). Not surprisingly, there were nearly twice as many significant comparisons when RNA was analyzed independently than when pooled (Figure 6.C), which may have resulted from the increased number of comparisons made when samples were considered independently. Interestingly, much less range in the simple fold changes was evident from sample pools. For instance, only one simple fold change exceeded 7.5, and only 3 ORFs yielded fold changes greater than 5-fold throughout the entire experiment. This result contrasts sharply with the much larger numbers of ORFs that exhibited wide fold changes when considered separately (see Table 1). Furthermore, when the RNA was pooled, many more differentially expressed ORFs were significant after mixed model analyses than with the $\log_2$ transformation tests. This result also contrasts with the results obtained when RNA was considered independently

in which roughly an equal number of ORFs were discovered for all fold change intervals (Table 1).

Biological trends in expression

As visualized in the overall LSM heat plot resulting from the mixed model analyses, expression of most ORFs within a steady state was relatively constant with time (Figure 7). Previous mixed model analyses have demonstrated that dye effects can be as significant, or more so, than treatment effects (10). Here, dye effects were important in many cases (4.2% of the total number of ORFs examined), but sampling time in the reactor was as important as dye in terms of assessing gene specific effects (Table 2). The effect due to sampling may be related to time since the culture began or it may be due to differences in RNA preparation or biological variability in gene expression. However, many of the ORFs showing a significant effect due to sampling time were previously shown to be induced in biofilm formation (23) and may be due to the formation of biofilms in the culture as time progressed. It is unlikely that the sampling effect would be due to significant evolution of the culture since selection was minimal in the vessel and the time course was short.

Table 3 lists selected ORFs that were up-regulated (adj. $P < 0.05$) at 85^{o}C or a slower growth rate as compared to growth at 80^{o}C, D=0.25 hr^{-1}. For all three fold change ranges tested (<1.25, <2.0, <5.0), increased transcription of genes encoding known heat shock proteins at the elevated temperature were better elucidated using a mixed model approach as compared to an analysis based on simple $\log_2$ data transformation. Of the most important genes determined from a previous study that examined the increased transcription of genes within 5 minutes following heat shock in batch culture (23), only the gene encoding the

GroEL protein was significantly up-regulated based on estimate differences from $\log_2$ transformations of the data. In contrast, mixed model analysis was able to capture the up-regulation in transcription of many known heat shock genes, in many cases even when the associated fold changes were <1.25. Up-regulation of genes encoding a putative ammonium transporter, various ribosomal proteins, and enzymes involved in the processing of the grown substrate (maltose) were up-regulated at a slower dilution rate and were better elucidated by using the mixed model in most cases. While these subtle growth-rate effects may seem counterintuitive, the synthesis of various ribosomal proteins also increased in *Escherichia coli* when the growth rate was reduced from 0.29 doublings/hr to 0.17 doublings/hr (14). Although the regulation of ribosomal gene expression in *E. coli* appears to occur at the translational level through an autogenous feedback mechanism (17), mechanisms of ribosomal gene expression regulation in *T. maritima* have not yet been examined.

The mixed model was also able to capture small (but biologically relevant) differences in transcription; this information would have been lost if small fold changes would have been discarded based on a cutoff even as high as '5-fold.' In fact, using a Bonferroni corrected P value of 0.05, the differential expression of a total of 885 comparisons (representing 231 different ORFs) were significant at fold changes less than 1.25, 2399 comparisons (323 ORFs) were significant at less than a 2-fold level, and 3923 comparisons (416 ORFs) at the 5-fold level.

By accounting for variation in gene expression during the progression of the culture through independently assessing biological replicates, it was possible to determine whether alterations in gene expression were due to the tested experimental conditions or to random fluctuations in expression within each steady state. Many ORFs showed a significant

difference in expression within a steady state (Figure 8). As shown, the difference in expression was not due to gene rank (based on average fluorescence intensities for each spot); many of the transcripts that were regulated most strongly occurred at low gene ranks and were not correlated with the average corrected fluorescence intensity for the given spot. Furthermore, allowing expression variance to be gene specific allows for clearer identification of responding ORFs that are statistically significant. As indicated in Figure 8, significantly fewer genes were differentially regulated by allowing variance to vary on a gene-by-gene basis, rather than held constant at a value equal to the median of the gene-specific variances, in the gene-specific mixed model analysis.

Clusters of orthologous genes (COGs) form a computationally-based framework for distinguishing functional categories of genes (30, 31). On a global scale, the three tested reactor states showed a broadly similar distribution of variances across COG functional categories (Figure 9). A notable exception is Category O (chaperones, protein turnover), which shows a variability in expression relative to the mean that is larger at the elevated temperature, possibly reflecting an unstable transcriptional response. Lipid metabolism (Category I) seems to be more variable when the cells are grown at steady state at 80°C than when the cells are growing more slowly or at an elevated temperature. While not investigated in detail here, it is possible that variation may result from influences such as differences in cellular microenvironments (e.g. nutrient and temperature gradients), differences between growth phase in different cells in the culture, phase variations, periods of rapid change in gene expression, genetic mutation and stochastic effects.

Figure 10 illustrates the collective transcriptional response to slower growth and higher temperature in terms of functional category (30, 31) designations. The most

pronounced effects at the slower growth rate occur from the up-regulation of genes involved in 4 functional categories: about 50% of the total differences between functional categories involve up-regulation at the slower growth rate. The transcription of genes from functional categories E (amino acid transport and metabolism), G (Carbohydrate transport and metabolism), K (transcription) and P (inorganic ion transport and metabolism) appear to be especially important during slower growth in the culture. Nearly one third of all the differences were genes involved in transport and metabolism of amino acids or carbohydrates; however, according to previous work elucidating the utilization of simple and complex sugars in *T. maritima* (4), there is good reason to believe that many of the amino acid transporters are incorrectly annotated and may transport sugars, as has been noted for ABC transporters in *Sulfolobus solfataricus* (5) and *Pyrococcus furiosus* ((13) and Shockley et al., personal communication). It is worthy of mention that the ORFs from all four of these functional categories showed more down-regulated genes than up-regulated genes for faster growth at the higher temperature. Transcription of genes from category T (signal transduction) contained much biological noise and was not considered further.

Genes from most functional categories are down-regulated during growth at higher temperature. While the genes encoding chaperones known to be important in heat shock response (e.g., GroEL and GroES, see Table 3) were up-regulated, an approximately equal number of genes involved in posttranslational modification, protein turnover, and chaperoning were up-regulated as were down-regulated. Roughly 75% of the genes displaying transcriptional differences due to temperature were down-regulated at the higher temperature. These ORFs belonged to categories E, G, K, P, and M (cell envelope biogenesis and lipid metabolism); no categories showed pronounced preferences to up-regulate the

expression of ORFs at higher temperature. The categories E, K, G, P may have been down-regulated as a result of growth at a faster specific growth rate, since they were up-regulated when grown more slowly.

Figure 11 summarizes the differential transcription events that were found to be significant through the mixed model analysis; a total of 375 transcripts were significantly changed during the course of the experiment. Of this number, 102 transcripts were differentially regulated at the elevated temperature and 318 transcripts were significantly different at the reduced dilution rate (specific growth rate). However, much of the variation between steady states was explained by variation within each steady state. As shown here, this suggests that genes that do not appear to be outliers in standard test statistics may have inherently lower variability in their expression levels than most of the other genes used in the analysis and should not be discounted based on only one experimental replicate. Of the 102 transcripts that were differentially regulated as the higher temperature, 23 ORFs (or 22.5%) were differentially expressed within a mechanical steady state. Also, of the 318 transcripts significantly changed at the lower growth rate, 71 ORFs (or 22.3%) were also changed within a steady state. This information would have been unavailable if the samples were not treated independently.

Concluding remarks

Many microarray analyses rely on naïve fold change criteria or associated test statistics that do not account for random or systematic errors, or effects arising from natural inflections in gene expression. We demonstrate here that mixed model analyses are able to accommodate many different levels of replication which incorporate multiple sources of

error and lead to improved biological inference. Here, the mixed model was applied to multiple samples generated from a single continuous culture run and used to detect subtle changes in growth rate and temperature that were not apparent from normalized simple fold change comparisons or differences arising from the associated $\log_2$ transformed data sets. A high agreement was apparent between simple fold changes and significance between differences of least squares estimates resulting from $\log_2$ transformations only for fold changes greater than 10-fold. However, mixed model analyses indicate that even high fold changes are artificially biased by effects of dye, array, or spot within array. In addition, it was demonstrated that data resulting from low fluorescence intensities may generate significant inferences if the associated errors are small. Finally, significant ORFs from experiments conducted by pooling RNA samples will allow the detection of significant differences in transcription abundance between compared samples, but such discovery may be artificially influenced by chance fluctuations in gene expression.

Supplementary material

The full set of differentially expressed genes will be published online at http:///www.che.ncsu.edu/extremophiles/ upon acceptance of manuscript.

Acknowledgements

This work was supported in part by grants from the Department of Energy (Energy Biosciences Program) and the National Science Foundation (Biotechnology Program). KRS acknowledges support from a Department of Education GAANN Fellowship. SBC acknowledges support from a NIEHS Traineeship in Bioinformatics.

References

1. **Auger, S., A. Danchin, and I. Martin-Verstraete.** 2002. Global expression profile of *Bacillus subtilis* grown in the presence of sulfate or methionine. J. Bacteriol. **184:**5179-5186.

2. **Bailey, J. E., and D. F. Ollis.** 1986. Biochemical engineering fundamentals. McGraw-Hill, Inc., New York, NY.

3. **Brown, S. H., and R. M. Kelly.** 1989. Cultivation techniques for hyperthermophilic archaeabacteria: continuous culture of *Pyrococcus furiosus* at temperatures near 100^{o}C. Appl. Environ. Microbiol. **55:**2086-2088.

4. **Chhabra, S. R., K. R. Shockley, S. B. Conners, K. L. Scott, R. D. Wolfinger, and R. M. Kelly.** 2003. Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima*. J. Biol. Chem. **278:**7540-7552.

5. **Elferink, M. G., S. V. Albers, W. N. Konings, and A. J. Driessen.** 2001. Sugar transport in Sulfolobus solfataricus is mediated by two families of binding protein-dependent ABC transporters. Mol. Microbiol. **39:**1494-1503.

6. **Hartzell, P. L., J. Millstein, and C. LaPaglia.** 1999. Biofilm formation in hyperthermophilic Archaea. Methods Enzymol. **310:**335-49.

7. **Hasseman, J.** 2001, posting date. TIGR Microarray Protocols. [Online.]

8. **Huber, R., T. A. Langworthy, H. Konig, M. Thomm, C. R. Woese, U. B. Sleytr, and K. O. Stetter.** 1986. *Thermotoga maritima* sp. nov. represent a new genus of unique extremely thermophilic eubacteria growing up to 90^{o}C. Arch. Microbiol. **144:**324-333.

9. **Hughes, T. R., M. J. Marton, A. R. Jones, C. J. Roberts, R. Stoughton, C. D. Armour, H. A. Bennett, E. Coffey, H. Dai, Y. D. He, M. J. Kidd, A. M. King, M. R. Meyer, D. Slade, P. Y. Lum, S. B. Stepaniants, D. D. Shoemaker, D. Gachotte, K. Chakraburtty, J. Simon, M. Bard, and S. H. Friend.** 2000. Functional discovery via a compendium of expression profiles. Cell **102:**109-126.

10. **Jin, W., R. M. Riley, R. D. Wolfinger, K. P. White, G. Passador-Gurgel, and G. Gibson.** 2001. The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. Nat. Genet. **29:**389-395.

11. **Kendziorski, C. M., Y. Zhang, H. Lan, and A. D. Attie.** 2003. The efficiency of pooling mRNA in microarray experiments. Biostatistics **4:**465-77.

12. **Kerr, M. K.** 2003. Design considerations for efficient and effective microarray studies. Biometrics **59:**822-828.

13. **Koning, S. M., M. G. Elferink, W. N. Konings, and A. J. Driessen.** 2001. Cellobiose uptake in the hyperthermophilic archaeon *Pyrococcus furiosus* is mediated by an inducible, high-affinity ABC transporter. J. Bacteriol. **183:**4979-4984.

14. **Milne, A. N., W. W. Mak, and J. T. Wong.** 1975. Variation of ribosomal proteins with bacterial growth rate. J. Bacteriol. **122:**89-92.

15. **Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, C. M. Fraser, and et al.** 1999. Evidence for lateral

gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature **399:**323-329.

16. **Newton, M. A., C. M. Kendziorski, C. S. Richmond, F. R. Blattner, and K. W. Tsui.** 2001. On differential variability of expression eatios: improving statistical inference about gene expression changes from microarray data. J. Comput. Biol. **8:**37-52.

17. **Nomura, M., R. Gourse, and G. Baughman.** 1984. Regulation of the synthesis of ribosomes and ribosomal components. Annu. Rev. Biochem. **53:**75-117.

18. **Oleksiak, M. F., G. A. Churchill, and D. L. Crawford.** 2002. Variation in gene expression within and among natural populations. Nat. Genet. **32:**261-266.

19. **Ozbudak, E. M., M. Thattai, I. Kurtser, A. D. Grossman, and A. van Oudenaarden.** 2002. Regulation of noise in the expression of a single gene. Nat. Genet. **31:**69-73.

20. **Peng, X., C. L. Wood, E. M. Blalock, K. C. Chen, P. W. Landfield, and A. J. Stromberg.** 2003. Statistical implications of pooling RNA samples for microarray experiments. BMC Bioinformatics **4**.

21. **Piper, M. D., P. Daran-Lapujade, C. Bro, B. Regenberg, S. Knudsen, J. Nielsen, and J. T. Pronk.** 2002. Reproducibility of oligonucleotide microarray transcriptome analyses. J. Biol. Chem. **277:**37001-37008.

22. **Pysz, M. A., K. D. Rinker, K. R. Shockley, and R. M. Kelly.** 2001. Continuous cultivation of hyperthermophiles. Methods Enzymol. **330:**31-40.

23. **Pysz, M. A., D. E. Ward, K. R. Shockley, C. I. Montero, S. B. Conners, M. R. Johnson, and R. M. Kelly.** 2004. Transcriptional analysis of dynamic heat-shock response by the hyperthermophilic bacterium *Thermotoga maritima*. Extremophiles.

24. **Quackenbush, J.** 2002. Microarray data normalization and transformation. Nat. Genet. **32 Suppl:**496-501.

25. **Rinker, K. D., C. J. Han, and R. M. Kelly.** 1999. Continuous culture as a tool for investigating the growth physiology of heterotrophic hyperthermophiles and extreme thermoacidophiles. J. Appl. Microbiol. Symp. Suppl. **85:**118S-127S.

26. **Rinker, K. D., and R. M. Kelly.** 2000. Effect of carbon and nitrogen sources on growth dynamics and exopolysaccharide production for the hyperthermophilic archaeon *Thermococcus litoralis* and bacterium *Thermotoga maritima*. Biotechnol. Bioeng. **69:**537-547.

27. **Schuler, M. L., and F. Kargi.** 1992. Bioprocess engineering. Prentice Hall, Englewood Cliffs, NJ.

28. **Schut, G. J., S. D. Brehm, S. Datta, and M. W. Adams.** 2003. Whole-genome DNA microarray analysis of a hyperthermophile and an archaeon: *Pyrococcus furiosus* grown on carbohydrates or peptides. J. Bacteriol. **185:**3935-3947.

29. **Shockley, K. R., D. E. Ward, S. R. Chhabra, S. B. Conners, C. I. Montero, and R. M. Kelly.** 2003. Heat shock response by the hyperthermophilic archaeon *Pyrococcus furiosus*. Appl. Environ. Microbiol. **69:**2365-2371.

30. **Tatusov, R. L., E. V. Koonin, and D. J. Lipman.** 1997. A genomic perspective on protein families. Science **278:**631-637.

31. **Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin.** 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29:**22-28.

32. **ter Linde, J. J., H. Liang, R. W. Davis, H. Y. Steensma, J. P. van Dijken, and J. T. Pronk.** 1999. Genome-wide transcriptional analysis of aerobic and anaerobic chemostat cultures of *Saccharomyces cerevisiae*. J. Bacteriol. **181:**7409-7413.

33. **Wittes, J., and H. P. Friedman.** 1999. Searching for evidence of altered gene expression: a comment on statistical analysis of microarray data. J. Natl. Cancer Inst. **91:**400-401.

34. **Wolfinger, R. D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules.** 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. **8:**625-637.

Tables

Table 1. Significant Changes in Gene Expression

simple fold change[∞]	number of ORFs with simple fold change	number of ORFs after log_2 transformation*	number of ORFs after mixed model*
<1.25	**1907**	**243**	**231**
1.25-1.5	**1905**	**214**	**219**
1.5-2.0	**1904**	**323**	**268**
2.0-3.0	**1897**	**359**	**271**
3.0-5.0	**1858**	**350**	**209**
5.0-10.0	**1018**	**265**	**115**
10.0-20.0	**68**	**57**	**23**
20.0-50.0	**41**	**40**	**6**
50.0-100.0	**24**	**23**	**3**
>100.0	**12**	**12**	**1**

*Shown is the number of unique, significant open reading frames differentially expressed (Bonferroni adjusted $P < 0.05$) for the given fold change interval. Bonferroni adjustments were made after the indicated normalization and were based on 148,368 total comparisons.

[∞]Simple fold change is calculated directly after total signal intensity normalization.

Note: ORFs do not sum over fold change intervals (1907 total unique ORFs), because the same ORF may be significant at multiple fold change intervals.

Table 2: Significance of Effects

	P<0.05	P<0.01
Treatment	**978 (51.3%)**	**674 (35.3%)**
Dye	**249 (13.1%)**	**81 (4.2%)**
Time	**246 (12.9%)**	**71 (3.7%)**
Treatment x Time	**70 (3.7%)**	**20 (1.0%)**

The given probabilities directly result from variance components from mixed model analyses and therefore do not have any Bonferroni correction. Percentages of significant ORFs of the total number examined (1907) are shown in brackets.

Table 3. Selected ORFs Up-regulated due to Temperature or Growth Rate Effects

Locus	Function	Fold < 1.25	$Hits_{LT}$	$Hits_{MM}$	Fold < 2.0	$Hits_{LT}$	$Hits_{MM}$	Fold < 5.0	$Hits_{LT}$	$Hits_{MM}$
Temperature Effects (85^oC vs. 80^oC)										
TM0373	dnaK protein	9	0	1	18	0	1	24	0	1
TM0374	Heat shock protein class I	4	0	4	16	0	16	24	0	24
TM0505	GroES	3	0	3	14	0	14	24	0	24
TM0506	GroEL	0	0	0	5	0	3	24	4	6
TM0816	Transcriptional regulator, MarR family	0	0	0	0	0	0	21	0	14
TM0851	Heat shock operon repressor HrcA	9	0	3	19	0	4	24	0	4
Growth Rate Effects ($D=0.17\ hr^{-1}$ vs. $D=0.25\ hr^{-1}$)										
TM0402	ammonium transporter	4	4	4	9	9	9	18	18	18
TM0403	nitrogen regulatory protein P-II	2	0	2	16	2	16	18	2	18
TM0451	ribosomal protein L33	4	1	1	13	1	4	18	1	6
TM0454	ribosomal protein L11	3	0	1	15	0	2	18	0	4
TM0455	ribosomal protein L1	2	2	0	6	6	0	6	6	0
TM0456	ribosomal protein L10	5	0	0	15	1	0	18	1	0
TM1835	cyclomaltodextrinase, putative	0	0	0	10	0	0	18	0	1
TM1840	Alpha-amylase	6	0	5	8	0	7	17	0	10
TM1845	pullulanase	1	0	0	8	0	0	18	0	1

[x]Listed are the number of times that selected ORFs were significantly up-regulated (Adj. $P < 0.05$) in either the $\log_2$ transformed ($Hits_{LT}$) or mixed model estimates ($Hits_{MM}$) at elevated temperature or decreased dilution rate for the indicated simple fold change ranges. For each ORF, there were a total of 24 comparisons made for each temperature effect and 18 comparisons to analyze each dilution rate effect.

Figures

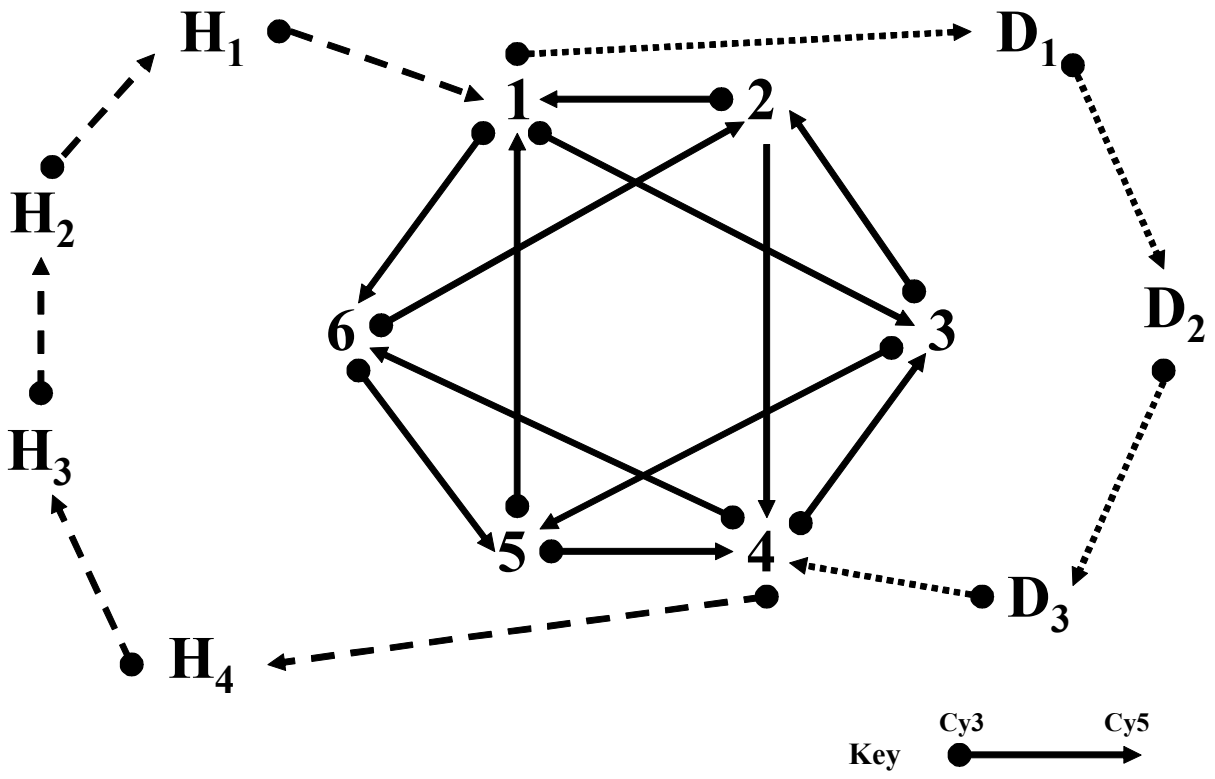
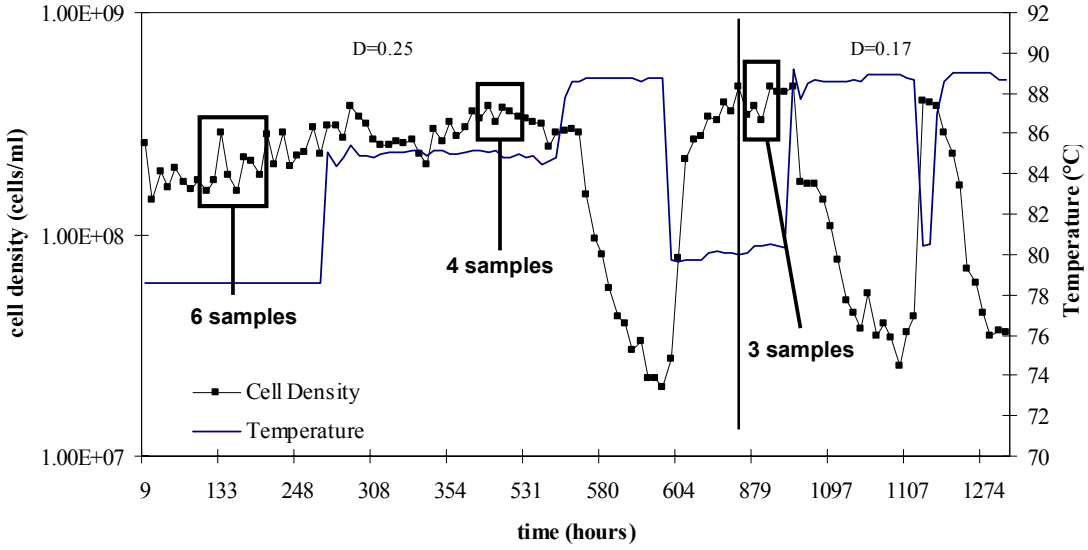


Figure 1. Loop design for the study of biological variability in *Thermotoga maritima.*
The arrow ends correspond to the Cy3 and Cy5 channels as follows: Cy3 •→Cy5.



Sample Harvesting Times (hours since inoculation)

Sample	T = 80°C, D=0.25 hr^{-1}	T = 85°C, D=0.25 hr^{-1}	T = 85°C, D=0.17 hr^{-1}
1	103.7	435.4	879.0
2	109.3	459.6	958.5
3	127.7	483.4	1000.0
4	133.4	501.7	•
5	151.8	•	•
6	157.5	•	•

Figure 2. Growth of *Thermotoga maritima* in a continuous culture. Six time points taken at 80°C, D=0.25 hr^{-1}; four time points taken at 85°C, D=0.25hr^{-1}; three samples taken at 80°C, D=0.17 hr^{-1}.

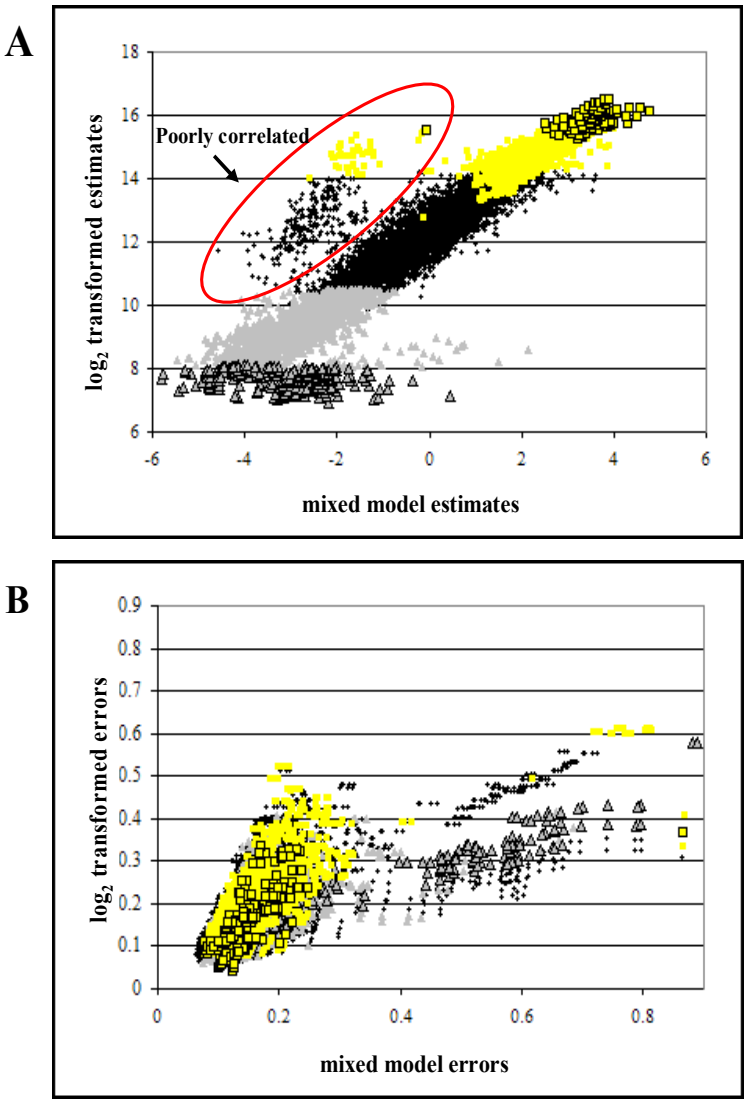


Figure 3. Least square means and standard errors after normalization. (A) Least squares means estimates of treatment effects and (B) standard errors of least squares means estimates are plotted for the $\log_2$ transformed and mixed model normalization analyses, respectively. For both instances, squares represent data points within the upper 10^{th} percentile of gene rank, diamonds the 10-90 percentile, and triangles the lower 10^{th} pencentile. Outlined squares and triangles represent the upper and lower 1 percentile of gene rank, respectively. Gene ranks were calculated as percentiles of average normalized fluorescence intensities within gene and treatment across dyes.

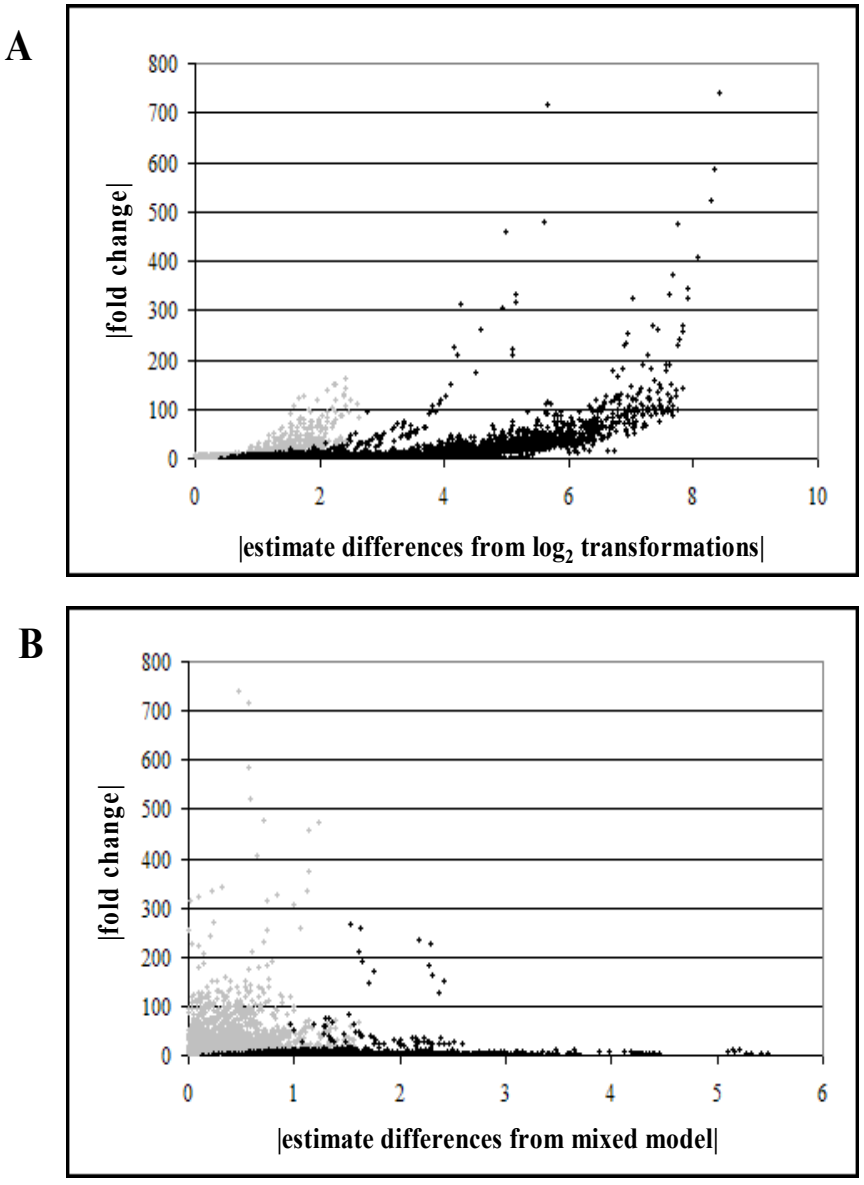


Figure 4. Simple fold changes as a function of differences in least square means estimates. Fold changes are shown as differences of least squares means estimates of treatments effects for (A) $\log_2$ transformed and (B) mixed model analyses. Points in black represent estimate differences with Adj. $P < 0.05$.

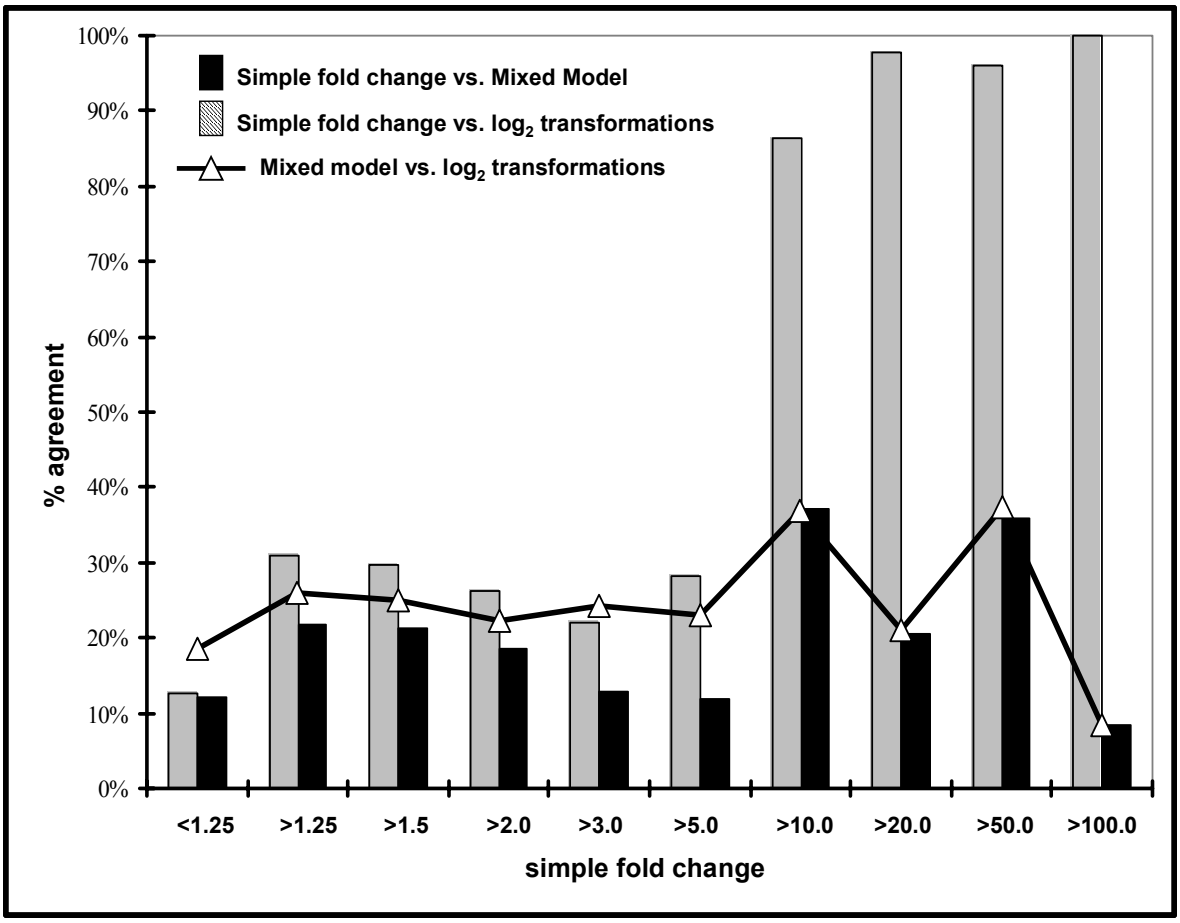


Figure 5. Agreement between simple fold change and normalization models. Agreement is based on the total number of significant ORFs (adjusted P<0.05) differentially regulated in the same direction between (i) simple fold change and $\log_2$ transformation model, (ii) simple fold change and mixed model and (iii) $\log_2$ transformation and mixed model, as indicated.

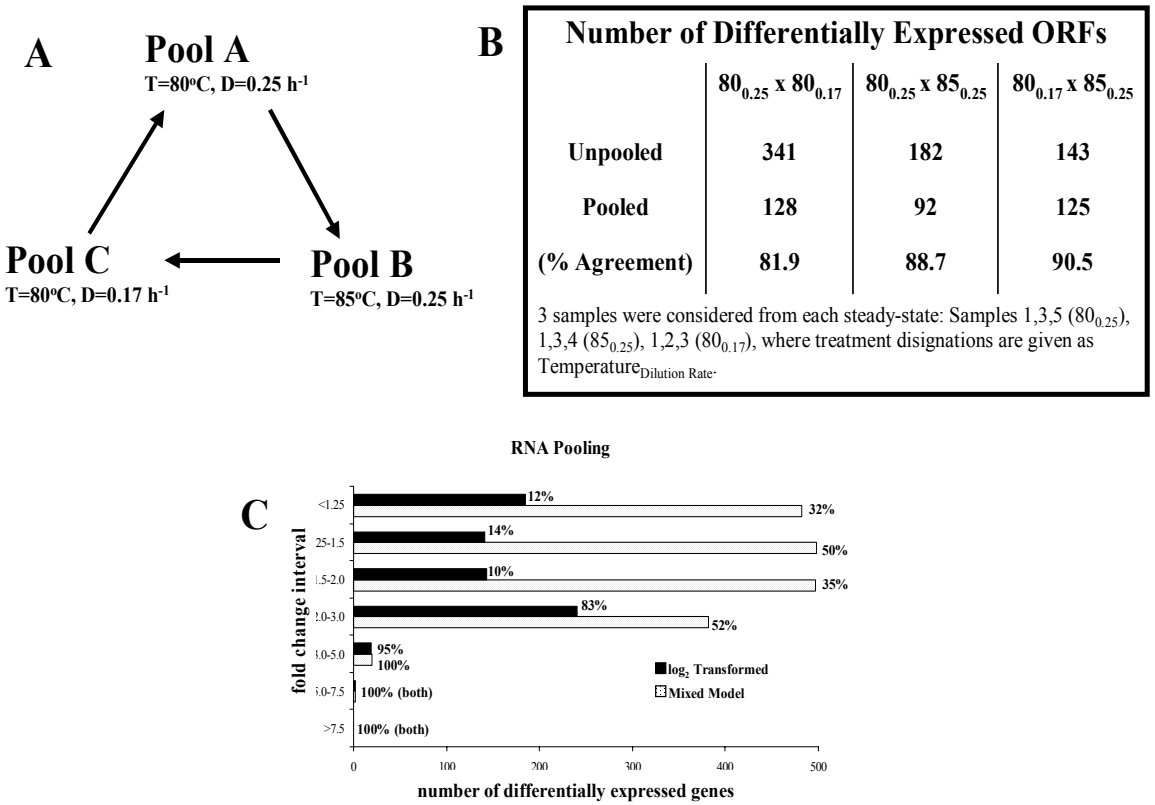


Figure 6. Effects of pooling RNA. (A) Three independently isolated and processed RNA samples were analyzed separately, and pooled and analyzed, according to the indicated loop design. Arrow ends correspond to the Cy3 and Cy5 channels as follows: Cy3 •→Cy5. (B) Percent agreement refers to the number of ORFs that are consistently differentially expressed or are not differentially expressed in the pooled and unpooled cases as determined by mixed model analysis. (C) Number of differentially expressed genes for indicated fold change interval resulting from pooled RNA samples. In all cases, significance of differential gene expression is according to an adjusted $P<0.05$).

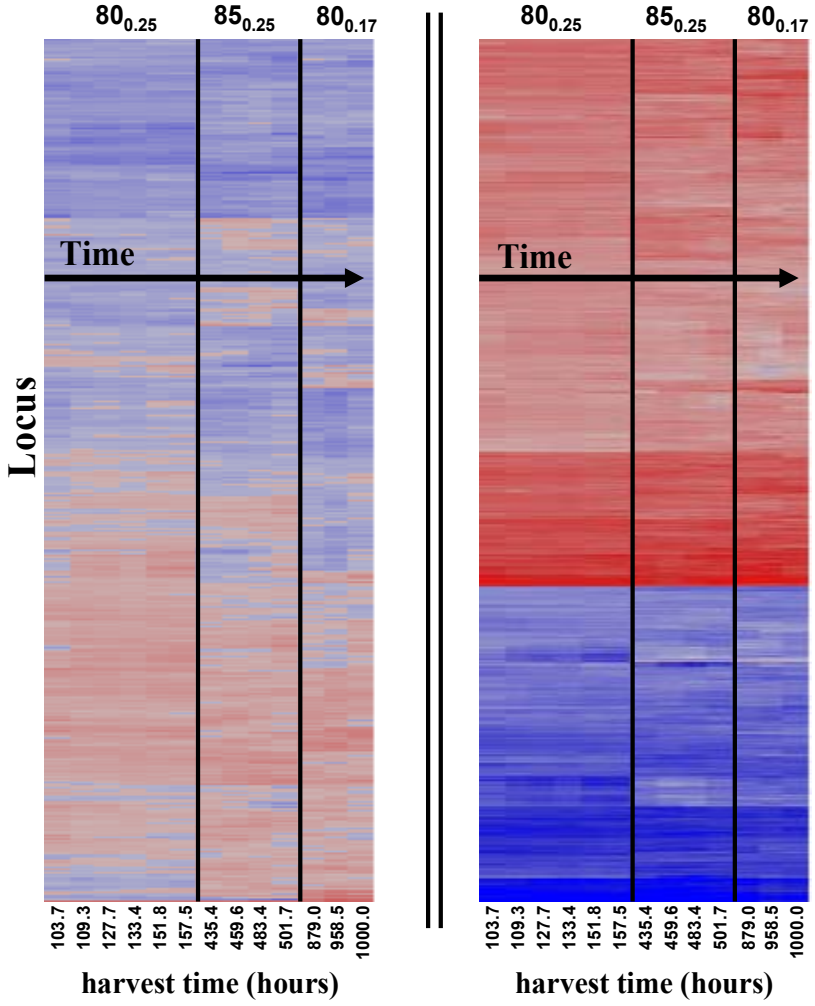


Figure 7. Hierarchical clusters constructed using least square means estimates. Harvest times (hours since reactor inoculation) and steady state regions are indicated. Each row represents a different gene in the organism. Treatment disignations are given as $Temperature_{Dilution\ Rate}$.

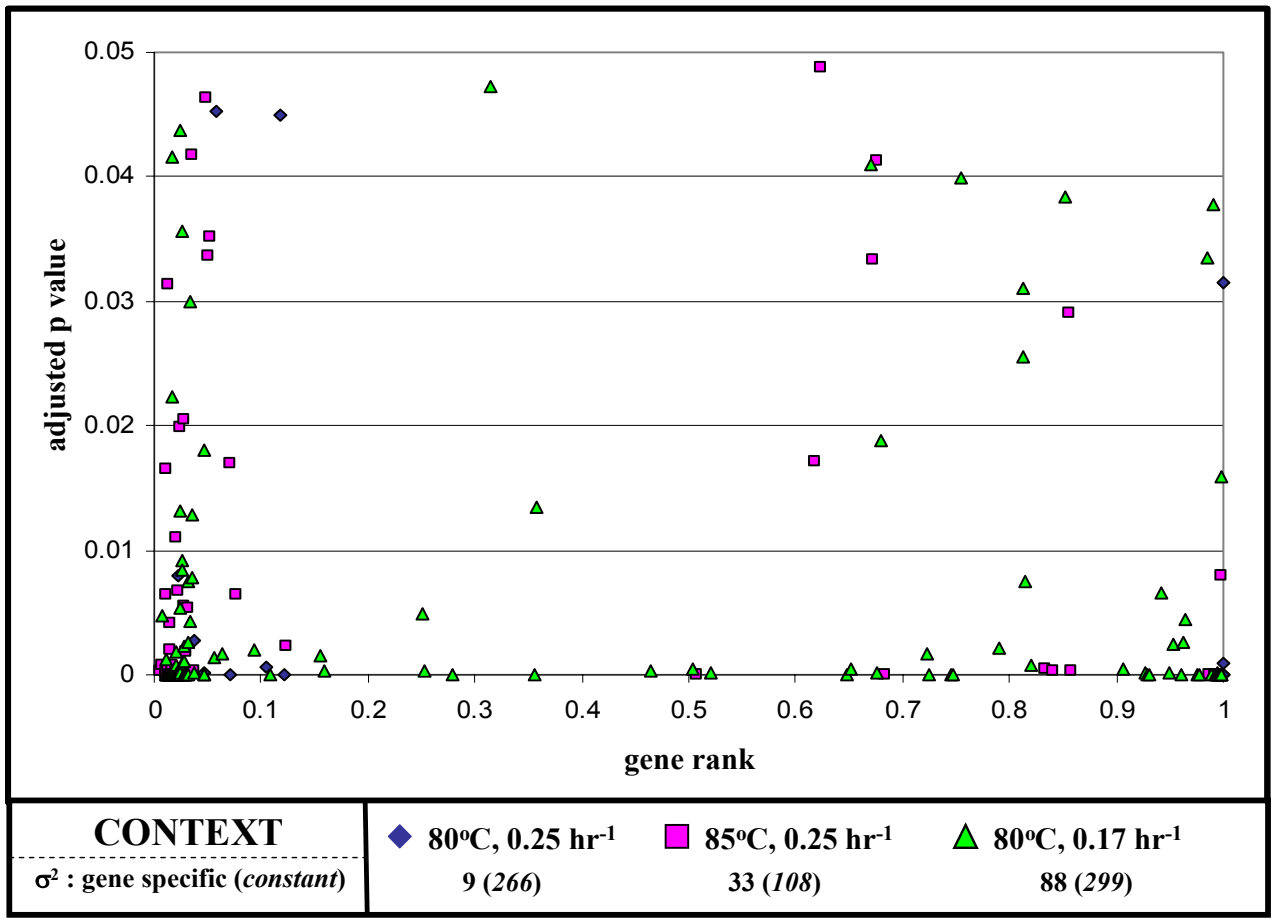


Figure 8. Adjusted P values and gene rank. Differentially expressed ORFs (adjusted P<0.05) within an indicated steady state are plotted against percentile of $\log_2$(fluorescence intensity) before $\log_2$ transformed or mixed model normalization.

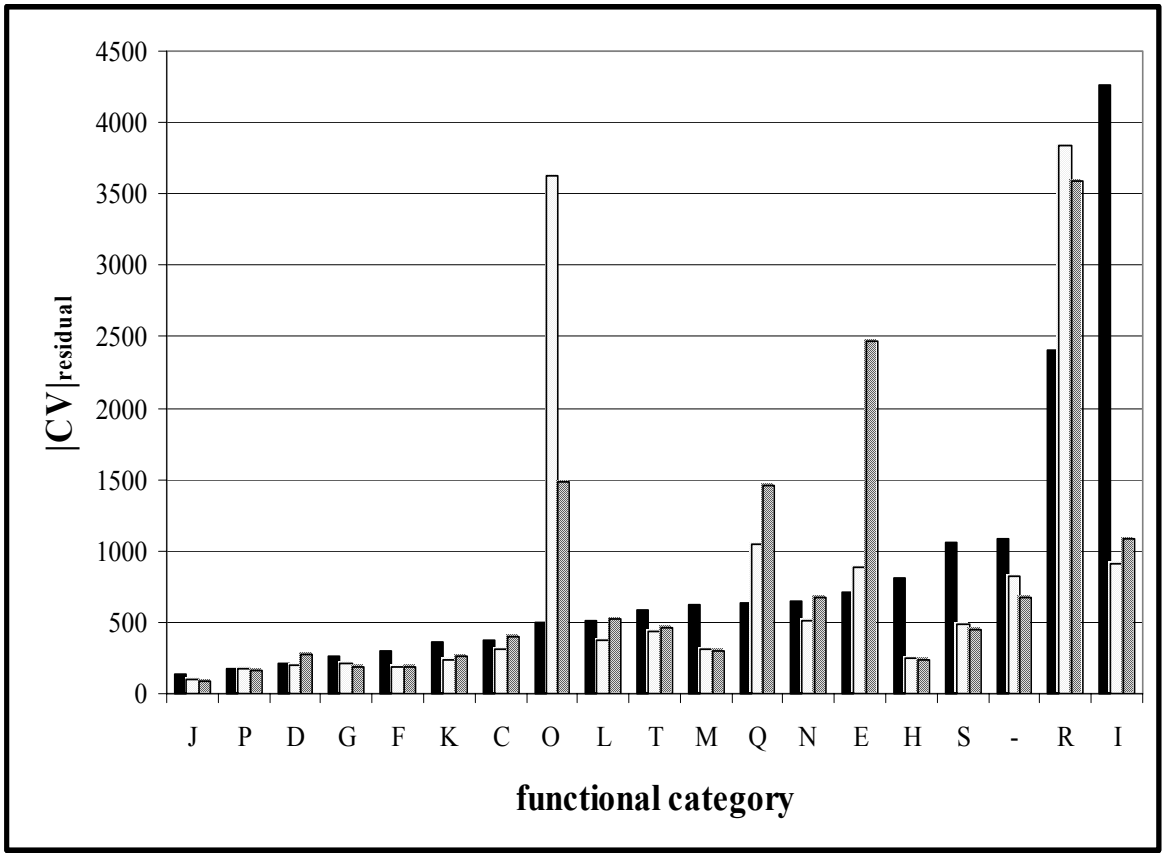


Figure 9. Transcriptional variation within steady states. Shown is variation in the absolute value of the coefficient of variation of residuals from the gene specific model within each of the three steady states. ■, 80°C, D=0.25 hr^{-1}; ⬚, 85°C, D=0.25 hr^{-1}; ▨, 80°C, D=0.17 hr^{-1}.

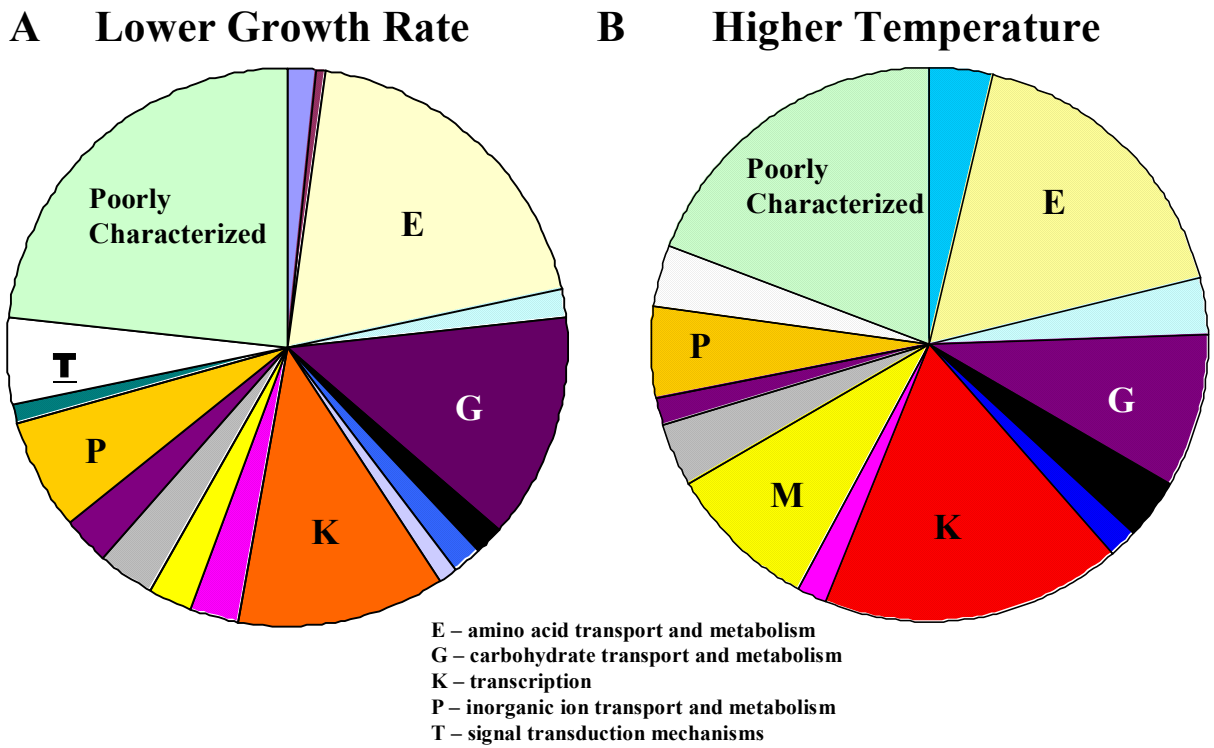


Figure 10. Distribution of differentially expressed ORFs among functional categories. Shown are differences in the number of significant ORFs (number of up-regulated ORFs minus number of down-regulated ORFs) between steady states (adj. $P<0.05$). Solid colors indicate that more genes were up-regulated, while dotted surfaces indicated that more genes were down-regulated, within a given functional category (30, 31).

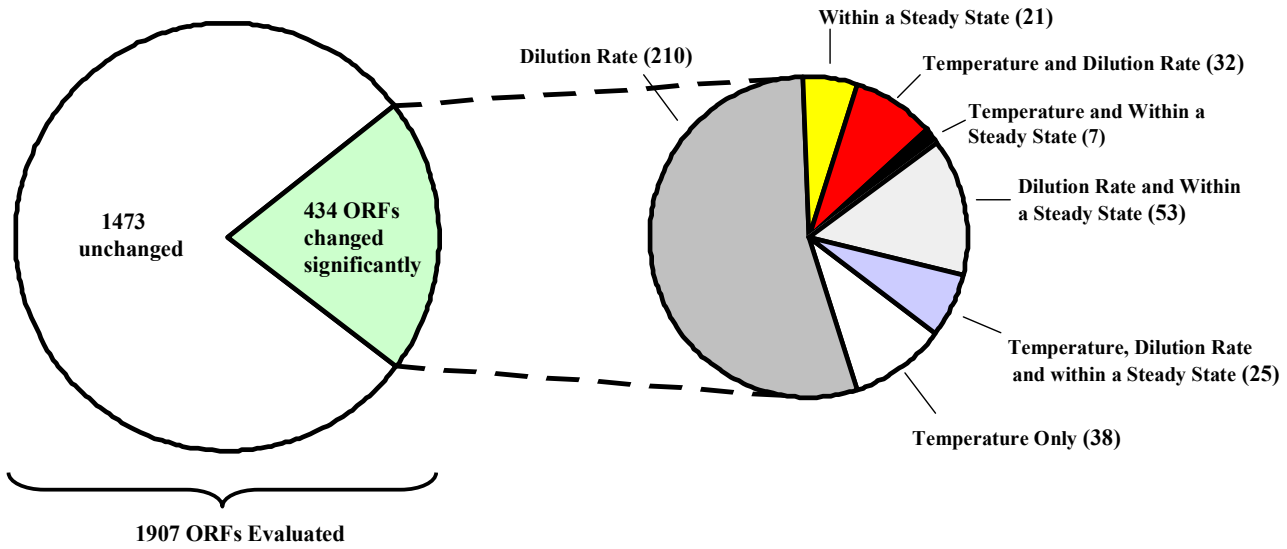


Figure 11. Summary of transcriptional responses within and between steady states. Significant findings (Adj. P<0.05) are reported for global transcriptional responses within and between the mechanical steady states at 80°C (D=0.25 hr^{-1} and D=0.17 hr^{-1}) and 85°C (D=0.25hr^{-1}). Differentially expressed ORFs within all three mechanical steady states are summed and reported as "Steady State." Significant ORFs in each pie chart are mutually exclusive to the stated category.

Chapter 4: Heat Shock Response by the Hyperthermophilic Archaeon *Pyrococcus furiosus*

Keith R. Shockley, Donald E. Ward[∞], Swapnil R. Chhabra[ß], Shannon B. Conners, Clemente I. Montero and Robert M. Kelly*

Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

[∞]Current address: Genencor International, Inc.
925 Page Mill Rd.
Palo Alto, CA 94304-1013

[ß]Current address: Biosystems Research Department
Sandia National Laboratories
Mailstop 9951, PO Box 969
Livermore, CA 94551-0969

Published in: ***Applied and Environmental Microbiology*** (April, 2003)
69: 2365-2371

Running Title: *Pyrococcus furiosus* heat shock

*Address inquiries to: **Robert M. Kelly**
Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

Phone: (919) 515-6396
Fax: (919) 515-3465
Email: rmkelly@eos.ncsu.edu

Abstract

Collective transcriptional analysis of heat shock response in the hyperthermophilic archaeon *Pyrococcus furiosus* was examined using a targeted cDNA microarray in conjunction with Northern analyses. Differential gene expression suggests that *P. furiosus* relies on a cooperative strategy of rescue (thermosome [Hsp60], small heat shock protein [Hsp20], two VAT-related chaperones), proteolysis (proteasome) and stabilization (compatible solute formation) to cope with polypeptide processing during thermal stress.

Introduction

Information gleaned from genome sequence data indicates that heat shock response in hyperthermophilic archaea has several distinguishing features. For example, hyperthermophilic archaea lack Hsp70 (DnaK), Hsp40 (DnaJ), and GrpE, all of which are centrally important in the heat shock response of most known microorganisms (26). Also, the major chaperonin found thus far in thermophilic archaea, an Hsp60 homolog referred to as the "thermosome," is more closely related to chaperonins associated with the eukaryotic cytosol (TriCC/CCT complex) than to the bacterial GroEL/ES system (19, 32). Energy-dependent proteolysis plays a major role during heat shock in bacteria in which genes encoding ATP-dependent proteases, such as *lon*, *clp*, and *hfl*, are linked to heat shock promoters (13). However, based on available genome sequence data, hyperthermophilic archaea lack the Clp and HflB (FtsH) family of proteins and have a different version of the Lon protease (43). Hyperthermophilic archaea, which typically have proteasomes, lack the eukaryotic ubiquitination pathway for selective protein degradation by the proteasome and, therefore, seem to modulate proteolysis at the protease level. Another interesting feature of hyperthermophilic archaeal heat shock response is the induced formation of unique compatible solutes that have been proposed to stabilize intracellular proteins against thermal denaturation (33). Whether compatible solutes reduce the need for protein turnover mechanisms is not known.

The relative contributions to the collective response of chaperones, chaperonins, proteases and compatible solutes during heat shock in hyperthermophilic archaea have yet to be examined. Here, the heat shock response of the hyperthermophilic archaeon *Pyrococcus furiosus* (9) was investigated using Northern analyses in conjunction with a targeted cDNA

microarray, based on genes encoding the thermosome, molecular chaperones, proteases, glycoside hydrolases and other relevant cellular functions expected to be affected during thermal stress.

Materials and methods

Experimental approach and data analysis

Relevant ORFs were located from the *P. furiosus* genome at NCBI (http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/framik?db=genome&gi=228) and BLAST searches of prokaryotic genomes found at the Institute for Genomic Research (www.TIGR.org). SCANPROSITE (http://expasy.cbr.nrc.ca/tools/scnpsit1.html) and PFAM HMM search tools (http://pfam.wustl.edu/hmmsearch.shtml) were used to verify the presence of putative catalytic domains. ORF fragments (generally 400 - 700bp) were selected from 201 different genes (about 10% of the genome) and were PCR-amplified and purified as reported previously (6) using *P. furiosus* genomic DNA. Purified PCR products were quantified (100 ng/□l), randomized and printed onto CMT-GAPS aminosilane-coated slides (Corning, Corning, NY) with a 417 Arrayer (Affymetrix, Santa Clara, CA). A total of 6 replicates of each ORF were printed onto each slide. DNA was attached to the substrate by UV crosslinking in at 250 mJ and baking at 75^{o}C for 2 hours. Five separate first-strand cDNA reactions (per sample, per slide) were prepared from *P. furiosus* total RNA using Stratascript (Stratagene) and random hexamers (Invitrogen Life Technologies, Carlsbad, CA) using indirect incorporation as previously described (15). The cDNA generated from each of the five reactions was pooled before hybridization to the slides; hybridizations and washes were performed as described in (16), except that no poly-T was added to the hybridization

mixture. Slides were scanned and signal intensity data was extracted using the Scanarray 4000 scanner and Quantarray software, respectively (GSI Lumonics, Billerica, MA). Local background intensity was subtracted from each spot signal, and spotting buffer (50% DMSO) was used to subtract global background. The signal from the Cy-3 channel was normalized to the signal from the Cy-5 channel based on total signal intensity.

P. furiosus DSM 3638 was grown on a sea salts-based medium (SSM) with 0.5% (w/v) tryptone + 0.2% (w/v) yeast extract as carbon and energy sources, as described previously (42), and monitored by cell enumeration as described elsewhere (6). RNA isolations were performed as described previously (6), except that RNA was further purified using the RNAqeuous™ RNA isolation kit (Ambion) after the first acid phenol extraction and subsequent ethanol precipitation (42). Northern analysis was carried out as described previously (6), with 20 μg of total RNA loaded in each lane. Cells were grown until mid-exponential phase at 90^{o}C on SSM and shifted to 105^{o}C for one hour; control cultures were allowed to continue growing at 90^{o}C for the same period of time (Figure 1A). The detection limit was determined by examining scatter plots of dye switch experiments. As shown in Figure 1B, differential dye incorporation did not produce an effect greater than 2-fold for any ORF on the array and is not significant for signal intensity values greater than 2000 units. Therefore, the results were not further data corrected for a dye effect. However, many genes were differentially expressed 2-fold or more with statistical significance (Figure 1C).

Three separate hybridization experiments were performed using the same RNA sample and the results were compared with data from an independent isolation and hybridization (Table 1). Negative controls (genes from mouse) were spotted onto the array to assess background noise. Ratios of microarray intensity data from the hybridization

experiments were combined and converted to a $\log_2$ scale; ORFs containing a negative ratio value (generally indicating very low transcript abundance) were discarded from further analysis. Genes that showed corrected intensity ratios of approximately 2-fold induction or repression ($\log_2 R_i > 1$ or $\log_2 R_i < -1$, where R_i is the intensity ratio) with a paired t-test (between unperturbed and perturbed growth) significance level of $p < 0.01$ (as shown in Table 1) were considered differentially expressed.

For complete information on signal intensity and fold changes for all genes included on the array, see our website (www.che.ncsu.edu/extremophiles/publications/Pfu_heatshock.html).

Results and discussion

Differential expression of genes during heat shock

Thermal stress was apparent by the effects on growth (Figure 1A) and the induction of known and putative stress genes present in *P. furiosus* (see Table 1). Therefore, the results given here are based on the combined effects of thermal stress and reduced cell growth that are collectively described by the term "heat shock." The genes encoding the major Hsp60-like chaperonin (thermosome) in *P. furiosus* (19) and the Hsp20-like small heat shock protein (23) were strongly induced as were two other molecular chaperones (VAT), belonging to the CDC48/p97 branch of the AAA^+ family. VAT is thought to participate in both protein unfolding (for proteolysis) and re-folding processes (12, 32). The *P. furiosus* genome encodes the α- and β-subunits of prefoldin, an ATP-independent chaperone found primarily in eukaryotes and archaea (25). The gene encoding the prefoldin β-subunit was down regulated upon heat shock, although the corresponding ORF was expressed at relatively high levels in both cases, which is consistent with reports suggesting that the genes encoding prefoldin are not induced by stress (27). The proposed protein folding cascade in *P. furiosus*, containing all known chaperone homologs in the organism, is shown in Figure 2.

Two damage repair proteins were monitored for response to heat shock. RadA and RadB in *P. furiosus* correspond functionally to RecA and Rad51 in bacteria and eukaryotes, respectively (20). The gene encoding RadA was elevated 2-fold during heat shock while RadB was unaffected. Although it was previously proposed that *rad*A was constitutively expressed in *P. furiosus* during a one-hour heat shock from 95°C to 108°C (20), the results here indicate that a heat shock-inducible DNA repair system is present in this organism. It is not known whether *P. furiosus* has an adaptive DNA repair system akin to the bacterial SOS

response, but here the gene encoding the *E. coli* DinF homolog in *P. furiosus* (4) was expressed at low levels under both conditions tested.

The proteasome appears be the only true ATP-dependent protease in *P. furiosus* (2). The only other ATP-dependent protease candidate encoded in the *P. furiosus* genome is an archaeal version of Lon that, unlike bacterial versions, lacks an ATP-binding domain sequence (10, 43). Here, the *P. furiosus lon* was not induced by heat shock, although this gene was expressed at relatively high levels under both stressed and unstressed conditions. Both proteasome β-subunits (β_1 and β_2) were induced somewhat upon heat shock (2-fold or less), while the expression of the α-subunit decreased 2-4 fold. The reasons for the decrease in α subunit gene expression of the *P. furiosus* proteasome upon heat shock is not known, although a similar observation has been made in mammalian cells (22). The conserved gene cluster containing the α subunit of the proteasome and the exosome (20) was differentially expressed in a concerted manner (data not shown). This result is consistent with predictions based on comparative genomic analyses (21, 29) and provides experimental support for the possible functional or physical coupling between selective protein degradation and RNA processing in archaea. The gene encoding PAN, the ATPase component of the 26S proteasome, was unaffected by heat shock.

The role of the proteasome in heat-shocked hyperthermophilic archaea and other prokaryotes is not known but presumably involves polypeptide processing. The thermophilic archaeon *Thermoplasma acidophilum* cannot grow without proteasome activity under heat shock conditions (31), although the proteasome is not essential during heat shock in the actinomycetes (8). The up-regulation of the β_1 subunit during heat shock is intriguing since this was found to be absent from the native 20S proteasome purified from *P. furiosus* (2).

Several ATP-independent proteases were also affected by heat shock (Table 1). Notably, the gene encoding pyrolysin, a membrane-associated protease with an endo-acting and subtilisin-like catalytic domain (41), was strongly repressed while a subtilisin-like protease (18) was strongly induced. Five peptidase-encoding genes were induced, including the gene encoding HtpX, which has been implicated elsewhere in surface protein expression related to changes in adhesiveness, cellular morphology and levels of surface-active antigens (38).

Hyperthermophiles accumulate compatible solutes during exposure to thermal stress (33). Levels of di-myo-inositol phosphate (DIP), the only reported temperature-dependent compatible solute in *P. furiosus*, were reported to increase as high as 20-fold when subjected to a temperature shift from 95^{o}C to 101^{o}C (28). In *Pyrococcus woesei*, DIP is presumed to be synthesized from glucose-6-phosphate by two enzymes: L-myo-inositol 1-phosphate synthase and a putative DIP-synthetase (34). Here, the gene encoding the former (PF1616) was strongly induced by thermal stress; the gene encoding the latter has yet to be identified. Another known compatible solute, trehalose, has been shown to stabilize proteins of various origins against thermal stress *in vitro* (17), but has not been reported to accumulate in *P. furiosus* under conditions of salt or thermal stress. Trehalose metabolism appears to be induced by heat shock in yeast (40) and *Salmonella enterica* (5). The closely related hyperthermophilic archaeon *Thermococcus litoralis* takes up trehalose through a high-affinity maltose/trehalose transport system (44). The genes encoding the putative trehalose synthase and MalE sugar binding protein were induced under heat shock (see Table 1), suggesting that *P. furiosus*, like *T. litoralis*, might take up trehalose available from yeast extract in its medium during stressed conditions or synthesize this compound (24).

Expression levels of glycoside hydrolase genes were, in general, very low during growth at 90°C on peptide-based medium (data not shown). However, there was a significant induction of these genes upon heat shock (see Table 1). The cellular motivation for expressing genes related to carbohydrate acquisition may relate to the increased demand in ATP during thermal stress, which could be met by increased glycolysis (30). However, the gene encoding glyceraldehyde-3-phosphate ferredoxin oxidoreductase (GAPOR) (37) was down-regulated significantly upon heat shock. It is also possible that futile cycles may be operational during heat shock response whereby the synthesis and catabolism of trehalose, glycerol, and glycogen help to control the energy balance of the cell, as observed in the stress response in *Saccharomyces cerevisiae* (1). Here, glycogen synthase and trehalose synthase gene expression levels (data not shown) were high during heat shock, which could trigger heightened glycoside hydrolase gene levels. Given the known capacity for *P. furiosus* and other hyperthermophiles to produce saccharide-based compatible solutes when subjected to stress (33), the acquisition of carbohydrates mediated by heightened levels of glycoside hyrolases for this purpose also needs to be considered. As such, the presence of small amounts of glycosides from yeast extract in the medium could have stimulated glycoside hydrolase gene expression under thermal stress.

Intensity comparisons from the Northern hybridizations on selected genes (Figure 3) were consistent with differential expression patterns observed in the microarrays. In addition, the microarray expression levels obtained here compared well with those reported by Schut et al. (35) who used a targeted microarray to study the effect of sulfur on maltose-grown *P. furiosus* at 95°C (see our website, http://www.che.ncsu.edu/extremophiles/publications/Pfu_heatshock.html, for details).

Comparative genomic analyses predicted conserved archaeal heat shock regulons consisting of genes encoding the small heat shock protein, a thermosome subunit, two VAT homolog CDC48-2 proteins, and two histones in *P. furiosus* (11). With the exception of the genes encoding the two histones, all of these genes in this experiment showed substantial increased fold changes and high levels of expression under heat shock conditions. While the gene encoding the archaeal histone A1 (PF1831) was not induced, it was expressed at very high levels under both normal and heat shock conditions.

Previous work in *Methanosarcina mazeii* with the *grpE-dnaK-dnaJ* chaperone gene cluster suggests that TBP (TATA-binding protein) and TFB (eukaryotic transcription factor IIB homolog) interact more strongly with stress-gene promoters during heat shock (7). These chaperone-encoding genes are not present in *P. furiosus,* but this organism's genome contains two different TFB-related genes. The presence of multiple TFB and TBP homologs in some species of archaea has led to speculation that these proteins may play a role similar to that of sigma factors in bacteria by recognizing promoters with different sequences (3). Indeed, one of the six TFB-related genes in *H. volcanii* is heat shock-induced (36). Northern analyses in this study reveal that PF0687 is induced while PF1377 expression levels are similar during unstressed growth and heat shock (see Figure 3).

Whether additional heat shock elements exist in *P. furiosus* and related hyperthermophilic archaea remains to be seen. At this point, heat shock response in this organism appears to involve a much more limited set of genes than is found for other prokaryotes examined to date. This is consistent with previous work with *Metallosphaera sedula*, an extremely thermoacidophilic archaeon (14). Further work is needed to understand heat shock gene regulation in *P. furiosus*, especially as it relates to less thermophilic

organisms. Efforts along these lines have recently been reported (39) and related studies will help elucidate thermal stress response mechanisms among the most thermally active microorganisms known. Finally, even though gene regulation is most often controlled at the level of transcription initiation, additional investigations are needed to confirm that the microarray data presented in this study conform to true biological response. Such efforts are currently underway in our laboratory.

Acknowledgements

This work was supported in part by grants from the Department of Energy (Energy Biosciences Program) and the National Science Foundation (Biotechnology Program). KRS acknowledges support from a Department of Education GAANN Fellowship. SBC acknowledges support from an IGERT Fellowship in Bioinformatics. The authors thank Bryon Sosinski and Len van Zyl for helpful discussions and Amy Grunden for providing selected proteolytic fermentation PCR products.

References

1. **Alexandre, H., V. Ansanay-Galeote, S. Dequin, and B. Blondin.** 2001. Global gene expression during short term ethanol stress in *Saccharomyces cerevisiae*. FEBS Lett. **498:**98-103.

2. **Bauer, M. W., S. H. Bauer, and R. M. Kelly.** 1997. Purification and characterization of a proteasome from the hyperthermophilic archaeon *Pyrococcus furiosus*. Appl. Environ. Microbiol. **63:**1160-1164.

3. **Bell, S. D., and S. P. Jackson.** 1998. Transcription and translation in archaea: a mosaic of eukaryl and bacterial features. Trends Microbiol. **6:**222-228.

4. **Bouyoub, A., G. Barbier, J. Querellou, and P. Forterre.** 1995. A putative SOS repair gene (dinF-like) in a hyperthermophilic archaeon. Gene **167:**147-149.

5. **Canovas, D., S. A. Fletcher, M. Hayashi, and L. N. Csonka.** 2001. Role of trehalose in growth at high temperature of *Salmonella enterica serovar typhimurium*. J. Bacteriol. **183:**3365-3371.

6. **Chhabra, S. R., K. R. Shockley, D. E. Ward, and R. M. Kelly.** 2002. Regulation of endo-acting glycosyl hydrolases in the hyperthermophilic bacterium *Thermotoga maritima* grown on glucan- and mannan-based polysaccharides. Appl. Environ. Microbiol. **68:**545-554.

7. **De Biase, A., A. J. L. Macario, and E. C. de Macario.** 2002. Effect of heat stress on promoter binding by transcription factors in the cytosol of the archaeon *Methanosarcina mazei*. Gene **282:**189-197.

8. **De Mot, R., I. Nagy, J. Walz, and W. Baumeister.** 1999. Proteasomes and other self-compartmentalizing proteases in prokaryotes. Trends Microbiol. **7:**88-92.

9. **Fiala, G., and K. O. Stetter.** 1986. *Pyrococcus furiosus* sp. nov represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100º C. Arch. Microbiol. **145:**56-61.

10. **Fukui, T., T. Egushi, H. Atomi, and T. Imanaka.** 2002. A membrane-bound archaeal Lon protease displays ATP independent proteolytic activity towards unfolded proteins and ATP dependent activity for folded proteins. J. Bacteriol. **184:**3689-3698.

11. **Gelfand, M. S., E. V. Koonin, and A. A. Mirinov.** 2000. Prediction of transcription regulatory sites in archaea by a comparative genomic approach. Nucleic Acids Res. **28:**695-705.

12. **Golbik, R., A. N. Lupas, K. K. Koretke, W. Baumeister, and J. Peters.** 1999. The janus face of the archaeal Cdc48/p97 homologue VAT: Protein folding versus unfolding. Biol. Chem. **380:**1049-1062.

13. **Gottesman, S.** 1996. Proteases and their targets in *Escherichia coli*. Annu. Rev. Genet. **30:**465-506.

14. **Han, C. J., S. H. Park, and R. M. Kelly.** 1997. Acquired thermotolerance and stressed-phase growth of the extremely thermoacidophilic archaeon *Metallospaera sedula* in continuous culture. Appl. Environ. Microbiol. **63:**2391-2396.

15. **Hasseman, J.** 2001, posting date. TIGR microarray protocols. [Online.] http://www.tigr.org/tdb/microarray/protocolsTIGR.shtml

16. **Hedge, P., R. Qi, R. Abernathy, C. Gay, S. Dharap, R. Gaspard, J. Earle-Hughes, E. Snesrud, N. Lee, and J. Quackenbush.** 2000. A concise guide to cDNA microarray analysis. Biotechniques **29:**548-562.

17. **Hottiger, T., C. De Virgilio, M. Hall, T. Boller, and A. Wiemken.** 1994. The role of trehalose synthesis for the acquisition of thermotolerance in yeast. II. Physiological

concentrations of trehalose increase the thermal stability of proteins in vitro. Eur. J. Biochem. **15:**187-193.

18. **Kannan, Y., Y. Koga, Y. Inoue, M. Haruki, M. Takagi, T. Imanaka, M. Morikawa, and S. Kanaya.** 2001. Active subtilisin-like protease from a hyperthermophilic archaeon in a form with a putative prosequence. Appl. Environ. Microbiol. **67:**2445-2452.

19. **Klumpp, M., and W. Baumeister.** 1998. The thermosome: archetype of group II chaperonins. FEBS Lett. **430:**73-77.

20. **Komori, K., T. Miyata, J. DiRuggiero, R. Holley-Shanks, I. Hayashi, I. K. Cann, K. Mayanagi, H. Shinagawa, and Y. Ishino.** 2000. Both RadA and RadB are involved in homologous recombination in *Pyrococcus furiosus*. J. Biol. Chem. **275:**33782-90.

21. **Koonin, E. V., Y. I. Wolf, and L. Aravind.** 2001. Prediction of the archaeal exosome and its connections with the proteasome and the translation and transcription machineries by a comparative-genomic approach. Genome Res. **11:**240-252.

22. **Kuckelkorn, U., C. Knuehl, B. Boes-Fabian, I. Drung, and P. M. Kloetzel.** 2000. The effect of heat shock on 20S/26S proteasomes. Biol Chem. **381:**1017-1023.

23. **Laksanalamai, P., D. L. Maeder, and F. T. Robb.** 2001. Regulation and mechanism of action of the small heat shock protein from the hyperthermophilic archaeon *Pyrococcus furiosus*. J. Bacteriol. **183:**5198-5202.

24. **Lamosa, P., L. O. Martins, M. S. Da Costa, and H. Santos.** 1998. Effects of temperature, salinity, and medium composition on compatible solute accumulation by *Thermococcus spp*. Appl. Environ. Microbiol. **64:**3591-3598.

25. **Leroux, M. R., M. Fandrich, D. Klunker, K. Siegers, A. N. Lupas, J. R. Brown, E. Schiebel, C. M. Dobson, and F. U. Hartl.** 1999. MtGimC, a novel archaeal chaperone related to the eukaryotic chaperonin cofactor GimC/prefoldin. Embo J. **18:**6730-6743.

26. **Macario, A. J. L., and E. C. de Macario.** 1999. The archaeal molecular chaperone machine: Peculiarities and paradoxes. Genetics **152:**1277-1283.

27. **Macario, A. J. L., and E. C. de Macario.** 2001. The molecular chaperone system and other anti-stress mechanisms in archaea. Front. Biosci. **6:**D262-D283.

28. **Martins, L. O., and H. Santos.** 1995. Accumulation of mannosylglycerate and di-myo-inositol-phosphate by *Pyrococcus furious* in response to salinity and temperature. Appl. Environ. Microbiol. **61:**3299-3303.

29. **Maupin-Furlow, J., S. Kaczowka, M. Ou, and H. Wilson.** 2001. Archaeal proteasomes: proteolytic nanocompartments of the cell. Adv. Appl. Microbiol. **50:**279-338.

30. **Nickells, R. W., and L. W. Browder.** 1988. A role for glyceraldehyde-3-phosphate dehydrogenase in the development of thermotolerance in *Xenopus laevis* embryos. J. Cell Biol. **107:**1901-1909.

31. **Ruepp, A., C. Eckerskorn, M. Bogyo, and W. Baumeister.** 1998. Proteasome function is dispensable under normal but not under heat shock conditions in *Thermoplasma acidophilum*. FEBS Lett. **425:**87-90.

32. **Ruepp, A., B. Rockel, I. Gutsche, W. Baumeister, and A. N. Lupas.** 2001. The chaperones of the archaeon *Thermoplasma acidophilum*. J. Struct. Biol. **135:**126-138.

33. **Santos, H., and M. S. da Costa.** 2002. Compatible solutes of organisms that live in hot saline environments. **4:**501-9.

34. **Scholz, S., S. Wolff, and R. Hensel.** 1998. The biosynthesis pathway of di-myo-inositol-1,1 '-phosphate in *Pyrococcus woesei*. FEMS Microbiol. Lett. **168:**37-42.

35. **Schut, G. J., J. Z. Zhou, and M. W. W. Adams.** 2001. DNA microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus*: Evidence for a new type of sulfur-reducing enzyme complex. J. Bacteriol. **183:**7027-7036.

36. **Thompson, D. K., J. R. Palmer, and a. C. J. Daniels.** 1999. Expression and heat-responsive regulation of a TFIIB homolog from the archaeon *Haloferax volcanii.* Mol. Microbiol. **33:**1081-1092.

37. **van der Oost, J., G. Schut, S. W. Kengen, W. R. Hagen, M. Thomm, and W. M. de Vos.** 1998. The ferredoxin-dependent conversion of glyceraldehyde-3-phosphate in the hyperthermophilic archaeon *Pyrococcus furiosus* represents a novel site of glycolytic regulation. J. Biol. Chem. **273:**28149-54.

38. **Vickerman, M. M., N. Mather, P. Minick, and C. Edwards.** 2002. Initial characterization of the *Streptococcus gordonii* htpX gene. Oral Microbiol. Immunol. **17:**22-31.

39. **Vierke, G., A. Engelmann, C. Hebbeln, and M. Thomm.** 2002. A novel archaeal transcriptional regulator of heat shock response. J Biol Chem. **278**:18-26

40. **Voit, E. O., and T. Radivoyevitch.** 2000. Biochemical systems analysis of genome-wide expression data. Bioinformatics **16:**1023-1037.

41. **Voorhorst, W. G. B., R. I. L. Eggen, A. C. M. Geerling, C. Platteeuw, R. J. Siezen, and W. M. deVos.** 1996. Isolation and characterization of the hyperthermostable serine protease, pyrolysin, and its gene from the hyperthermophilic archaeon *Pyrococcus furiosus*. J. Biol. Chem. **271:**20426-20431.

42. **Ward, D. E., S. W. M. Kengen, J. van der Oost, and W. M. de Vos.** 2000. Purification and characterization of the alanine aminotransferase from the hyperthermophilic archaeon *Pyrococcus furiosus* and its role in alanine production. J. Bacteriol. **182:**2559-2566.

43. **Ward, D. E., K. R. Shockley, L. S. Chang, R. D. Levy, J. K. Michel, S. B. Conners, and R. M. and Kelly.** 2002. Proteolysis in hyperthermophilic microorganisms. Archaea **1:**63-74.

44. **Xavier, K. B., L. O. Martins, R. Peist, M. Kossmann, W. Boos, and H. Santos.** 1996. High-affinity maltose/trehalose transport system in the hyperthermophilic archaeon *Thermococcus litoralis*. J. Bacteriol. **178:**4773-4777.

Tables

Table 1. Differential Expression of Selected ORFs

Annotation	locus	$log_2 \pm$ STDEV (1st Repl.)	Fold	$log_2 \pm$ STDEV (2nd Repl.)	Fold
Chaperone-related genes					
Small heat shock protein (class I)	PF1883	2.95 ± 0.34	>7.7	2.80 ± 0.12	>6.9
VAT homolog	PF1882	2.78 ± 0.18	6.9	1.89 ± 0.04	3.7
VAT homolog	PF0963	2.21 ± 0.27	4.6	1.45 ± 0.05	2.7
Thermosome, single subunit	PF1974	2.00 ± 0.25	>4.0	2.07 ± 0.16	>4.2
Prefoldin homolog (beta subunit)	PF0380	-0.76 ± 0.19	-1.70	-1.02 ± 0.03	-2.0
ATP-dependent proteases					
Proteasome, subunit beta (PsmB-1)	PF1404	1.03 ± 0.33	2.0	0.98 ± 0.05	2.0
Proteasome, subunit beta (PsmB-2)	PF0159	0.89 ± 0.21	1.9	0.47 ± 0.05	1.4
ATP-dependent Regulatory Subunit (PAN)	PF0115	0.25 ± 0.16	1.2	-0.12 ± 0.05	-1.1
ATP-dependent LA (Lon) *	PF0467	-1.53 ± 0.24	-2.9	-0.38 ± 0.14	>-1.3
Proteasome, subunit alpha (PsmA)	PF1571	-1.96 ± 0.28	-3.9	-1.16 ± 0.03	-2.2
ATP-independent proteases/peptidases					
Subtilisin-like protease	PF0688	2.81 ± 0.31	7.0	4.04 ± 0.07	16.5
ArgE/peptidase	PF1185	2.18 ± 0.22	4.5	1.89 ± 0.05	3.7
HtpX heat shock protein	PF1597	2.16 ± 0.10	4.5	1.60 ± 0.04	3.0
Similar to endo-1,4-beta-glucanase (ytoP)	PF1861	1.67 ± 0.23	3.2	1.21 ± 0.07	2.3
D-aminopeptidase	PF1924	1.52 ± 0.18	2.9	1.96 ± 0.05	3.9
Signal sequence peptidase I, SEC11	PF0313	1.10 ± 0.23+	2.1	1.43 ± 0.05	2.7
Methionine aminopeptidase (MAP) (Pep M)	PF0541	-1.00 ± 0.15	-2.0	0.41 ± 0.03	1.3
Putative proline dipeptidase	PF0747	-1.18 ± 0.21	-2.3	0.32 ± 0.02	1.3
Heat shock protein X	PF1597	-1.18 ± 0.31	-2.3	-0.58 ± 0.05	-1.5
Carboxypeptidase I	PF0456	-1.27 ± 0.19	-2.4	-2.26 ± 0.03	-4.8
Pyrolysin	PF0287	-2.29 ± 0.19	>-4.9	-1.53 ± 0.05	-2.9
Glycoside hydrolases					
Beta-glucosidase (Cel1A)	PF0073	4.18 ± 0.20	18.1	3.25 ± 0.09	9.5
Putative methyltransferase (deacetylase)	PF0137	3.47 ± 0.28	11.1	1.83 ± 0.01	3.6
Endo-beta-1,3-glucanase (Lam16)	PF0076	2.67 ± 1.13	6.4	1.93 ± 0.15	3.8
Chitinase (Chi18A)	PF1234	2.14 ± 0.29	4.4	2.62 ± 0.11	6.2
Beta-galactosidase precursor Gal35 (put.)	PF0363	1.72 ± 0.21	3.3	1.85 ± 0.03	3.6
Alpha amylase (Amy57)	PF0272	1.64 ± 0.56	3.1	-1.34 ± 0.07	-2.5
Beta-mannosidase (Man1)	PF1208	1.46 ± 0.16	2.8	1.25 ± 0.03	2.4
Put. alpha-dextrin endo-1,6-α-glucosidase	PF1108	1.34 ± 0.27	2.5	1.59 ± 0.05	3.0
Chitinase (Chi18B)	PF1233	1.26 ± 0.67	2.4	2.87 ± 0.14	7.3
Beta-glucosidase (Cel1B)	PF0442	1.17 ± 0.23	2.3	0.39 ± 0.05	1.3

Table 1 (cont.)

Other					
Trehalose/maltose binding protein (*malE*)	PF1739	2.69 ± 0.17	6.5	1.29 ± 0.06	2.4
Spermidine synthase	PF0127	1.41 ± 0.28	2.7	1.99 ± 0.02	4.0
Recombinase, radA	PF1926	1.39 ± 0.20	>2.6	1.02 ± 0.04	2.0
Putative sugar binding protein (*malE-like*)	PF1938	1.34 ± 0.19	2.5	-1.48 ± 0.21	-2.8
Putative trehalose synthase	PF1742	1.29 ± 0.31	2.4	0.78 ± 0.08	1.7
Recombinase, radB	PF0021	0.46 ± 0.32	1.4	-0.24 ± 0.07	-1.2
Damage-inducible protein (dinF homolog)	PF1850	0.08 ± 0.23^{+}	1.1	-0.46 ± 0.06	-1.4

*Archaeal Lon proteins are missing an ATP-binding domain (43). $^{+}p<0.02$ (all other reported observations for $p<0.01$)

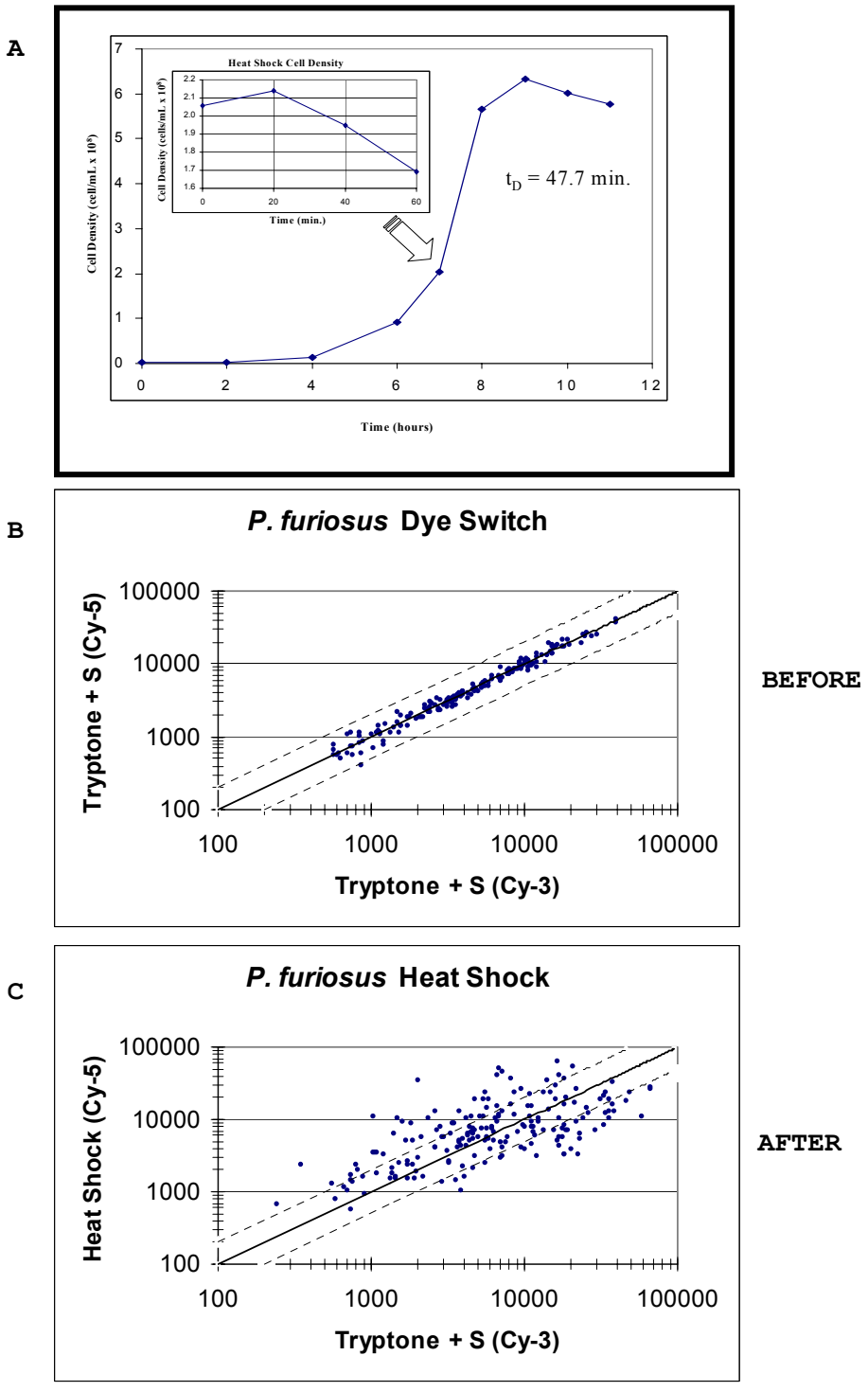


Figure 1. (A) Growth curve for *P. furiosus* grown on tryptone + S^0 (1% w/v) in sea salts medium at 90°C. Cells were subjected to a 60 min. heat shock at 105°C during exponential growth. (B) Reciprocally labeled mRNA from cells grown at 90°C shows consistent fluorescent signal intensities (SI). (C) Signal intensities of heat shock vs. unperturbed growth. Upper and lower diagonal lines indicate two-fold differential expression.

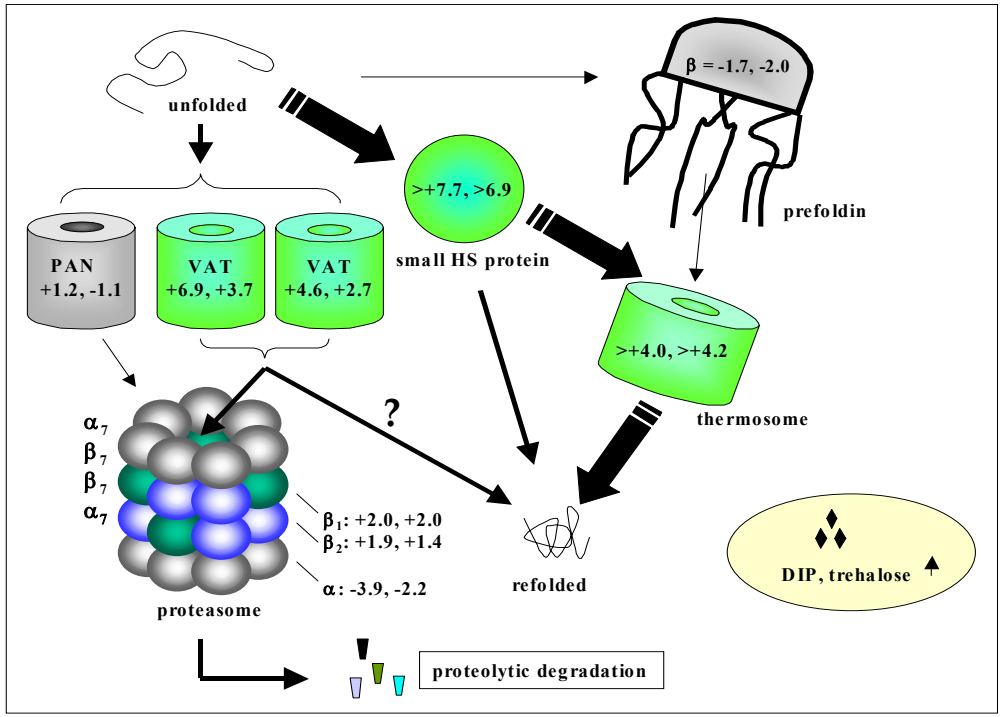


Figure 2: Protein folding cascade in *Pyrococcus furiosus* based on differential gene expression of ORFs. Fold values are presented for the heat shock experiment and the biological replicate. *P. furiosus* appears to utilize primarily Hsp20 and Hsp60 (and possibly VAT) as the major components of the refolding cascade while relying on then proteasome for energy-dependent proteolysis.

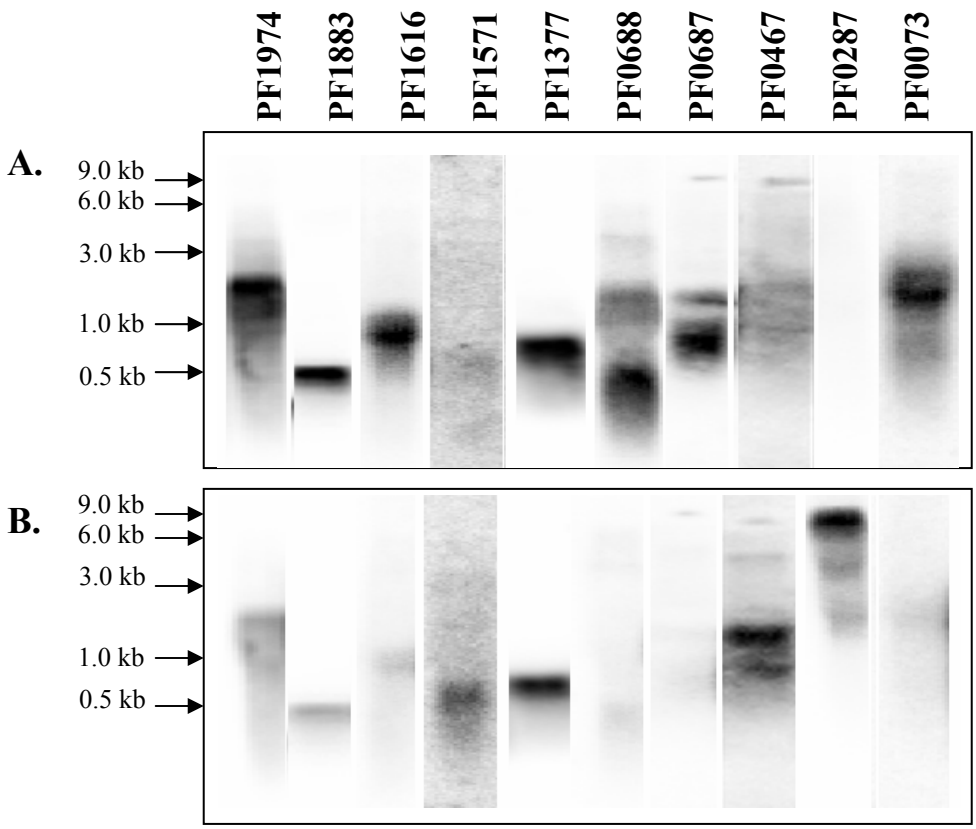


Figure 3. Northern analyses of selected genes in the *P. furiosus* experiment for 105°C (A) and 90°C (B) conditions.

Locus	Annotation	Fold change	Predicted transcript size (bp)
PF1974	thermosome - Hsp60	> 4.0 (> 4.2)	1650
PF1883	small HS protein - Hsp20	> 7.7 (>6.9)	513
PF1616	myo-inositol-1-phosphate synthase	NI	1152
PF1571	proteosome α-subunit	-3.9 (-2.2)	783
PF1377	TFIIB homolog (TFB)	NI	903
PF0688	subtilisin-like protease	7.0 (16.5)	594
PF0687	TFIIB homolog (TFB)	NI	852
PF0467	Lon	-2.9 (> -1.3)	3141
PF0287	pyrolysin	> -4.9 (-2.9)	4239
PF0073	β-glucosidase (Cel1)	18.1 (9.5)	1476

Chapter 5: Comparative Transcriptional Analyses of the Heat shock Response in Hyperthermophiles Using the Model Archaeon *Pyrococcus furiosus* and the Model Bacterium *Thermotoga maritima*

Keith R. Shockley, Shannon B. Conners, Matthew R. Johnson, Clemente I. Montero, Stephanie L. Bridger and Robert M. Kelly*

Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

To Be Submitted to: ***Journal of Bacteriology*** (June, 2004)

Running Title: Heat Shock in *Pyrococcus furiosus*

*Address inquiries to: **Robert M. Kelly**
Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

Phone: (919) 515-6396
Fax: (919) 515-3465
Email: rmkelly@eos.ncsu.edu

Abstract

Differential gene expression during heat shock was followed over the course of 90 minutes using full genome cDNA microarrays (>99% of the ORFs) corresponding to open reading frames in the hyperthermophilic archaeon *Pyrococcus furiosus* and the hyperthermophilic bacterium *Thermotoga maritima*. Global changes in both organisms were found to follow a transient expression pattern with only nine genes (including 5 hypothetical proteins and the small heat shock protein) differentially expressed throughout the first hour of stress in *P. furiosus*; 111 genes were differentially regulated in *T. maritima* for the full 90 minutes of stress. Approximately 10% of the *P. furiosus* genes were immediately up- or down-regulated both in media containing (1) maltose and yeast extract and (2) tryptone and elemental sulfur, while over 37% of the genes in this organism significantly changed in only one of the tested media ($P<0.001$). Even though the organisms in this study come from two separate domains of life, they employ similar strategies for dealing with heat shock at the onset of thermal stress, including the early induction of genes encoding chaperones and proteins involved in DNA repair and replication, and consistently decreased transcription of genes involved in energy expenditure. In contrast to transcriptional analyses of mesophilic organisms, neither of these hyperthermophilic species showed significant induction of genes related to ATP-dependent proteolysis. Furthermore, the induction of regulators throughout the time course and the consistent repression of genes involved in energy metabolism (including the modified Embden-Myerhoff, Entner-Douderoff, and pentose phosphate pathways and gluconeogenesis) indicate that long term survival strategies for heat shock exist in these organisms. Overall, the results presented here reveal much similarity and difference

between how hyperthermophilic microorganisms from different domains adapt their transcriptomes to cope with a sudden temperature upshift.

Introduction

Complete genome sequence information provides an opportunity to evaluate the relationship between genetic inventories and transcriptional phenotypes responsible for coping with stress across a very wide range of prokaryotic life. It is now possible not only to decipher the presence or absence of stress response proteins encoded in microbial genomes, but also to discover how organisms which lack these important molecules adapt to abrupt changes in temperature. It is currently known that, despite differences in transcription and regulation strategies, many molecular chaperones involved in stress response are highly conserved across domains of life (5) and are intricately linked to numerous basic functional processes inside of cells (8). Furthermore, the major heat shock proteins are vital to proper cell function under both normal and stressed conditions (13, 34) which has lead to increasing interest in the identification and regulation of the key components responding to heat-induced damage.

Temperature-induced response mechanisms in the mesophilic bacteria *Escherichia coli* and *Bacillus subtilus* are understood much better than the corresponding mechanisms in more distantly related bacteria or the archaea. Bacteria and Archaea have much different transcriptional machinery and regulatory strategies. The bacterial RNA polymerase can recognize sigma factors bound to distinct promoter elements (15, 29) while the basal transcription apparatus in the archaea, as homologues of components found in the eukarya (1, 15, 32), require the *in vivo* binding of TBP and TFB proteins to promoter regions before transcription can begin (27). In *E. coli* and *B. subtilis*, it is well known that important components of the heat shock response include genes involved in protein folding and turnover (7, 18, 20, 23, 35). However, many of the genes and regulatory elements central to

these processes are limited to a relatively narrow phylogenetic distance and are not found in all hyperthermophiles. The hyperthermophilic bacteria and some members of the archaea contain some or all of the familiar heat shock genes (hrcA, dnaK, grpE, groEL, groES, dnaJ), but these genes are absent in the hyperthermophilic archaea. Furthermore, the interplay of multiple sigma factors and the entire acid shock response are not readily identifiable in the available hyperthermophilic genome sequences (13). This suggests that the functions served by known systems in the proteobacteria may be performed by unrelated or distantly related sets of proteins in organisms adapted to growth above 80^{o}C. Many genes involved in the heat shock response continue to be discovered as it is becoming clearer that growth environment (14, 22) and the dynamics of transcription (28) influence how cells respond to the heat-induced damage occurring at elevated temperatures.

Here, full genome DNA microarrays were used to investigate heat shock response over time in two model heterotrophic hyperthermophiles; the euryarchaeon *Pyrococcus furiosus* and the bacterium *Thermotoga maritima*. Due to the absence of suitable genetic systems for hyperthermophiles, DNA microarrays have emerged as important means to investigate gene regulation in these species (19, 24-26). Dynamic information focusing on the development of the heat shock response in these organisms over the course of a 90-minute exposure to a large temperature upshift was used to examine the instant and longer-term survival strategies in these organisms. Furthermore, to investigate the impact of growth medium on transcriptional response, *P. furiosus* was grown in the presence of maltose and yeast extract as well as proteolytically (tryptone and elemental sulfur).

Materials and methods

Primer design and PCR

Open reading frames of all known genes in the annotated genome sequences of *Pyrococcus furiosus* DSM 3638 and *Thermotoga maritima* MSB8 were located at the site for The Institute for Genomic Research at (http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=ntpf01) and (http://www.tigr.org/tigrscripts/CMR2/GenomePage3.spl?database=btm), respectively. DNA primers were designed with similar annealing temperatures and minimal hairpin formation using Genomax 2.0 (Informax, Bethesda, MD). The probes were PCR-amplified in the PTC-100 Thermocycler (MJ Research, Inc., Waltham, MA) using Taq polymerase (Boehringer, Indianapolis, IN) and genomic DNA isolated from cell cultures according to previously published protocols (21). After amplification, the PCR products were quantitated and purified to a concentration of 100 ng/μl using QIAquick PCR purification kits (Qiagen, Valencia, CA). Purified PCR products were re-suspended in 50% DMSO, randomized, dispensed evenly into microarray printing plates (Genetix) and printed onto ULTRAGAPS aminosilane-coated microscope slides (Corning, Corning, NY) with a QArray-Mini Arrayer (Genetix, London, UK). The DNA was attached to the substrate by UV crosslinking in a GS GeneLinker UV Chamber (BioRad, Hercules, CA) at 250 mJ and then baked at 75^{o}C for 2 hours.

Growth of microorganisms

Thermotoga maritima (DSM 3109) and *Pyrococcus furiosus* were cultured anaerobically in a 14.0 L fermentor (New Brunswick Scientific, Edison, NJ) on Sea Salts

Medium (SSM) using either tryptone + S^0 (without yeast extract) or maltose as a carbon/energy source, as described previously (19). *T. maritima* was grown at 80°C until early log phase. After a 500 mL sample was collected, the temperature set point was shifted to 90°C (it took approximately 2 minutes to reach 90°C). *P. furiosus* was grown at 90°C until early log phase. After an 800mL sample was collected, the temperature set point was shifted to 105°C (it took approximately 4 minutes to reach 105°C). In all cases, samples were taken at 0, 5, 30, 60 and 90 minutes after reaching 90°C. Approximately 20 ml of culture were collected prior to sampling at each time point to eliminate pre-existing fluid in the sampling lines. At each time point, 800 ml of culture were withdrawn and immediately put on ice until processed for RNA extraction (see below). One ml of sample was removed for cell density enumeration by epifluorescent microscopy with acridine orange stain (9).

RNA isolation

P. furiosus was grown until mid-exponential phase at 90°C on SSM in a 14.0L fermentor and then immediately subjected to a heat shock at 105°C. To harvest, the cells were passed through a coffee filter to remove residual elemental sulfur and centrifuged at 7500 rpm (8145g) for 22 min. at 4°C. Cells were then re-suspended in ice-cold SSM, aliquotted into 2.0 mL eppendorf tubes (Ambion) and centrifuged at 13,800g for 30 sec. The resulting pellet was re-suspended in 85 µL of cold TE buffer, and then ruptured in 625 µL of lysis buffer (50 mM glucose, 10 mM EDTA, 25 mM Tris). The lysate was passed through a 20-gauge needle to shear the genomic DNA, after which 62.5 mL of 2 M sodium acetate (pH 5.2) was added to the resulting mixture. Then, an equal volume of acidic phenol/chloroform (5:1) was added, the aqueous phase was extracted, and the RNA was ethanol-precipitated

overnight at –20°C. After washing with 70% ethanol, the RNA pellet was re-suspended in 10 mM Tris (pH 8.5) and passed through fiber filter columns from the RNAqeuous™ RNA isolation kit (Ambion). Integrity of the RNA was confirmed by visual inspection on 1.0% native agarose gels as well as by measuring the A_{260}/A_{280} ratio with a DU® 640 Spectrophotomer (Beckman Coulter, Inc., Fullerton, CA).

Preparation of cDNA and hybridization

First-strand cDNA was prepared from total RNA using Stratascript™ (Stratagene) and random hexamer primers (Invitrogen Life Technologies, Carlsbad, CA) by the incorporation of 5-[3-Aminoallyl]-2'-deoxyuridine-5'-triphosphate (aa-dUTP) (Sigma) according to (6). The generated cDNA products were purified using the QIAquick PCR purification kit and reacted with monoreactive Cy-3 and Cy-5 NHS-esters (Amersham Biosciences, Inc., Piscataway, NJ). Another round of purification was used to remove unincorporated dyes. The hybridizations and washes were performed as described by (6) except that no poly-T was added to the hybridization mixture. The slides were scanned using the Scanarray™ 4000 scanner (GSI Lumonics, Canada Billerica, MA) found in the NCSU Genome Research Lab. Signal intensity data for all experiments was extracted using Quantarray (GSI Lumonics).

Mixed model analyses

Replication of treatments, arrays, dyes, and cDNA spots allowed the use of analysis of variance (ANOVA) models for data analysis (33). Loop designs were constructed as indicated in the text and reciprocal labeling was utilized for all samples so that dye effects could be estimated. Spot intensities obtained from Scanarray (Perkin Elmer) and imported

directly into SAS (SAS Institute, Cary, NC). A linear normalization ANOVA model (33) was used to estimate global variation in the form of fixed (dye (D), treatment (T)) and random (array (A), block (B) and spot (S)) effects and random error using the model $\log_2(y_{ijklmn}) = A_i + D_j + T_k + A_i(B_lS_m) + A_i(B_l) + \varepsilon_{ijklm}$. A gene-specific ANOVA model was then used to partition the remaining variation into gene-specific effects using the model $r_{ijklmn} = \mu + A_i + D_j + T_k + A_i(B_lS_m) + A_i(B_l) + \gamma_{ijklm}$.

For complete information on significance of expression changes, fold changes, pairwise volcano plots, and hierarchical clustering for all of the genes included on the array, see our website (follow the microarray link, data to be posted upon acceptance of the manuscript) at: http:///www.che.ncsu.edu/extremophiles/.

Results and discussion

Experimental approach and general results

Full genome cDNA microarrays for *P. furiosus* and *T. maritima* were constructed that included >99% of the predicted genes in the genomes of each of these organisms. Of interest here was to follow the time-dependent transcription of these genes after exposure to supraoptimal temperatures (a shift from 90^{o}C to 105^{o}C for *P. furious*; a shift from 80^{o}C to 90^{o}C for *T. maritima*). *P. furiosus* and *T. maritima* were grown using maltose and yeast extract as carbon/energy sources. In order to assess the effect of growth environment, *P. furiosus* was also grown using peptide fermentation in the presence of elemental sulfur during a temperature shift from 90°C to 105°C. Though many ORFs in its genome encode putative oligopeptide transporter genes, *T. maritima* is unable to grow proteolytically. Previous work has determined that many of the annotated oligopeptide transporters are expressed when *T. maritima* is grown on sugars ((2) and Conners et al., personal communication). A loop design was constructed for each of the experiments described above (Figure 1). In order to assess statistical significance of the results, it was necessary to establish appropriate statistical threshold for differential expression. Here, a threshold of $P<0.001$ captured the immediate induction of important molecular chaperones involved in heat shock response in previous studies (19, 26). Figure 2 shows growth of both organisms after the application of heat shock under the conditions examined. As shown in Figure 3, gene expression induction due to heat shock is broadly similar by functional category (30, 31) in both *P. furiosus* and *T. maritima*.

A total of 2016 different ORFs were spotted onto the *P. furiosus* array; 1594 ORFs had at least one significant change in expression ($P<0.001$) between the unstressed baseline

and a subsequent time point (0, 5, 30, 60 or 90 min.) after thermal stress in either growth on maltose or tryptone. A total of 1252 ORFs were differentially expressed on tryptone and sulfur without yeast extract in the medium while 1089 ORFs were differentially expressed when grown on maltose and yeast extract. The same maltose-based medium was used to cultivate *T. maritima*; a total of 1908 ORFs were printed onto the *T. maritima* arrays, but only 1033 had at least one significant comparison for growth on maltose. As shown in Table 1, many ORFs were induced immediately after heat shock was applied in both organisms.

Many more ORFs were consistently differentially expressed in *T. maritima* than were differentially expressed in *P. furiosus* (Tables 2 and 3). For instance, a total of 2 ORFs were up-regulated for the full 90-minute heat shock (PF0624 and PF0952) and 4 ORFs were down-regulated the full 90 minutes of thermal stress (PF0043, PF0152, PF1786 and PF1800) in *P. furiosus*. The expression pattern of PF0043 (phosphoenolpyruvate synthase) indicates that gluconeogenesis may be down-regulated throughout heat shock in *P. furiosus*. In *T. maritima*, 55 ORFs were up-regulated and 56 ORFs were down-regulated throughout the 90 minutes after heat shock (43 of these were hypothetical proteins). Figure 4 summarizes the gene expression differences between heat shock time points (0, 5, 30, 60, 90 min.) and unstressed growth for *P. furiosus* and *T. maritima*. Clusters indicate the most remarkable differences in gene expression for both organisms.

Heat shock response in Pyrococcus furiosus

Table 4 shows important features of the heat shock response in *P. furiosus*. A significance threshold of $P < 0.01$ was used for purposes of comparison with a previous study (26) in which *P. furiosus* was grown on tryptone + S^0 for one hour and differential gene

expression was measured on a targeted microarray. Although much agreement was found between the current study and the previous study, there are notable exceptions. Subunits encoding the proteasome, trehalose/maltose binding protein, HtpX heat shock protein and spermidine synthase were differentially regulated in the previous analysis but did not show differential gene expression in this analysis at the $P<0.01$ level. These discrepancies might be due to differences in media formulation (the current study utilized growth on tryptone without yeast extract while the previous study incorporated yeast extract into the medium), differences in thermal heat transfer (the current study utilized a well-mixed fermentor instead of performing experiments in bottles) or data analysis. Here, mixed model analysis was utilized in order to parse effects due to dye, array, and spot from treatment effects, while such modeling was absent from the earlier analysis. Nevertheless, Table 4 indicates the thermosome, the small heat shock protein and VAT homologs were induced in both experiments (also see Figure 5).

For *P. furiosus*, there are 4 notable clusters which show dramatic differences between the baseline and later time points (see Figure 4). The clusters indicated by $C_{P1.a}$ and $C_{P1.b}$ in Figure 5 show ORFs that are dramatically up-regulated in *P. furiosus* under proteolytic- and sugar-based growth. The fact that *P. furiosus* cells were experiencing thermal stress was confirmed by the early induction of several known and putative stress genes (see Figure 5). As found after one hour of thermal stress in a previous study of *P. furiosus* involving growth on tryptone, elemental sulfur and yeast extract (26), the genes encoding the thermosome (PF1974, the major Hsp60-like chaperonin in *P. furiosus*), the small heat shock protein (PF1982) and an AAA+ ATPase (PF1883) were strongly up-regulated immediately upon heat shock.

Previous work with hyperthermophiles suggests that DNA denaturation does not damage sequence integrity as long as DNA molecules are covalently closed (4). Indeed, it has been demonstrated that there is no correlation between GC content of DNA and thermophily (4). Even so, temperature-induced damage could cause depurination and cytosine deamination reactions which could lead to mismatches in DNA sequence, requiring DNA repair to ensue (16). In addition, increased temperature leads to double helix unwinding (3), which may increase the need for DNA stabilization processes in these organisms (e.g., supercoiling) to maintain appropriate topology of the DNA molecule. Interestingly, many of the genes initially up-regulated upon thermal stress are involved in putative DNA repair or folding processes in $C_{P1.a}$ and $C_{P1.b}$ (Figure 5). These included genes encoding two transposases (PF2024, PF1918), two subunits of a topoisomerase (PF1578, PF1579), a DNA invertase (PF2023), a RecA homolog (PF1931), radA (PF1926), and two genes encoding proteins related to DNA repair (PF2020 and PF2018). Previous work has indicated that *Pyrococcus* species have genes that encode enzymes that are important in DNA repair mechanisms in heat shock response, including RadA, which has DNA-dependent ATPase, DNA pairing, and strand exchange activities (11). Although it was proposed that *rad*A was constitutively expressed in *P. furiosus* at 95°C (11), the gene encoding radA was up-regulated in the first 5 minutes on both tryptone and maltose growth and radB (PF0021) was immediately up-regulated for growth on tryptone with $P<0.001$ (data not shown). Taken together, the results here for a shift from 90-105°C suggest that a heat shock-inducible DNA repair system is present in this organism, consistent with the discovery of a similar system in *Pyrococcus abyss*i which responds to temperature and radiation-induced DNA damage (10). Whether *P. furiosus* has an adaptive DNA repair system akin to the bacterial SOS response is

not known, although here the gene encoding the *E. coli* DinF homolog in *P. furiosus* (PF1850) was not differentially expressed within the first 5 minutes of heat shock.

Cluster C_{P4} in Figure 5 shows that many large- and small-subunit ribosomal proteins are induced during heat shock in *P. furiosus*. The dramatic induction of the genes encoding these ribosomal proteins is probably related to a temperature-induced increased growth rate in these organisms at the onset of thermal stress. After the first 5 minutes of heat shock, the transcription of the ORFs encoding these ribosomal proteins decreases rather sharply, probably reflecting diminished growth of the cells in the culture.

Figure 6 illustrates the global heat shock response as differences in the numbers of ORFs differentially expressed ($P<0.001$) over relevant functional categories (30, 31) as a function of time after heat shock. As shown in Figure 6, many genes involved in categories K (transcription), L (DNA replication, recombination and repair), and T (signal transduction) are up-regulated early after heat shock while many ORFs in categories C (energy production and conservation), J (translation, ribosomal structure and biogenesis) and O (Posttranslational modification, protein turnover, chaperones) are down-regulated late in heat shock response.

Effect of medium composition on heat shock response in P. furiosus

Figure 7 shows a growth substrate dependent effect on differential gene expression in *P. furiosus*. Cluster C_{P2} indicates that many ORFs which were strongly down-regulated during heat shock response for growth on maltose were up-regulated when grown on tryptone and sulfur. Included in this number are two ORFs (PF1935, amylopullalanase; PF1934, hypothetical protein) from the maltodextrin utilization operon (12) and one ORF (PF1743, hypothetical protein) from the maltose utilization operon. The open reading frame

corresponding to PF1934 shows 58% identity with a domain from *Thermococcus hydrothermalis* pullulanase and is probably related to the utilization of maltodextrin. This indicates that *P. furiosus* immediately down-regulates the expression of ORFs involved in utilizing maltose when grown on maltose, but induces these ORFs upon heat shock when grown on tryptone. One gene in this group encodes a VAT homolog (PF0963), and two ORFs (PF1432 and PF1430) encode subunits of an NADH dehydrogenase subunit.

Cluster C_{P4} from Figure 7 shows ORFs that were up-regulated for growth on maltose but down-regulated for growth on tryptone upon heat shock. Genes in this group include an aldehyde:ferredoxin oxidoreductase (P0346) and dipeptide transport system permease (PF1409) that are probably utilized during proteolytic growth. Other genes related to substrate processing include genes encoding an NDP-sugar synthase (PF0868), hydrogenases (PF1328, PF1911), NADH dehydrogenase (PF1445), a putative ABC transporter ATP-binding protein (PF1238), and a 4-aminobutyrate aminotransferase.

Heat shock response in Thermotoga maritima

Figure 8 includes clusters that yield significant differential gene expression of ORFs in *T. maritima* upon heat shock from 80°C to 90°C. Cluster C_{T3} reveals the immediate induction of the major heat shock genes *hrcA-grpE-dna*J (TM0851-TM0850-TM0849), *groES-groEL* (TM0505-TM0506), and *dnaK-sHSP* (TM0373-TM0374) after a 10°C increase, as has been found previously (19). As was the case with *P. furiosus* under both growth conditions, genes encoding ribosomal subunits increased immediately upon heat shock in clusters C_{T3} and C_{T2}. Most of these subunits decreased dramatically after the first 5 minutes of thermal stress, probably reflecting changes in growth rate in the culture. Cluster

$C_{T3.a}$ shows ORFs that were immediately up-regulated and were maintained at high expression levels throughout the course of thermal stress monitored on this array. Most of these genes encoded hypothetical proteins, although genes encoding a sugar ATP-binding protein (TM1276) and thiH protein (TM1267) showed increased levels of gene expression.

While Figure 8 shows genes that were immediately induced upon thermal stress, Clusters C_{T1} and $C_{T3.b}$ indicated that a latent gene expression induction pattern was present in the heat shock of *T. maritima*. In cluster C_{T1}, genes encoding two putative alpha-glucosidases (TM0434 and TM1068) and a transcriptional regulator (TM1069) were up-regulated after 90 minutes. Similarly, as found previously (19), a transcriptional regulator from the Mar family of regulators (TM0816) and a transcriptional regulator from the TetR family (TM0823) were up-regulated late in heat shock.

Comparison of heat shock response between P. furiosus and T. maritima

Pyrococcus furiosus is a member of the Euryarchaeota from the archaeal domain of life, while *Thermotoga maritima* is from the bacterial domain of life. While these two organisms originate from different domains, homologous ORFs with very high sequence identities found within in their genome sequences reveal that a substantial amount of lateral gene transfer events has taken place between domains. This observation is intriguing give the widely divergent transcriptional machinery of archaea and bacteria. In order to discover conservation of common heat shock strategies between the two domains of hyperthermophiles, both genomes were searched for pairwise sequence homologs containing at least 40% similarity between *P. furiosus* and *T. maritima*. This search revealed a total of 560 ORFs with a pairwise best similarity between the two organisms, over 25% of all ORFs

in each genome. However, of this total, only 16 ORFs responded to heat shock in a concerted fashion (see Figure 9). Most of these 16 ORFs encode ribosomal proteins that are located very close to each other in the genome. The total includes 11 ORFs that encode ribosomal proteins, 2 conserved hypothetical proteins, a translation elongation factor and two heat shock genes from the *hsp60* (groEL/thermosome) and *hsp20* (small heat shock protein) families.

While previous analyses of *T. maritima* and *P. furiosus* heat shock used targeted microarrays (19, 26), the present study encapsulates a full-genome analysis of heat shock response. Here, it is possible to monitor the transcriptional response of every known and putative chaperone in the *P. furiosus* genome (17). As shown in Table 5, *P. furiosus* contains known and putative chaperones for which transcriptional response data has never been reported. Of this total, only a possible archaeal BAG-1 chaperone was immediately induced under heat shock under both growth conditions (Table 2). A PPIase (PF1401), known to be induced under heat shock conditions, was induced under proteolytic growth and a possible archaeal Hip homolog (PF0335) identified in a previous search for eukaryal-like chaperones in archaea (17) was induced under growth of maltose and yeast extract. Here, it was found that the homolog to β-subunit of prefoldin, an ATP-independent chaperone, was down-regulated under heat shock, which was confirmed in a previous study (26).

It is believed that ABC transporters are the predominant means for transporting sugars, di/oligopeptides, metals, compatible solutes and other important substrates into hyperthermophilic cells from the environment. In addition, some ABC transporters may be used to export signal peptides for quorum sensing, balance pH, or remove toxic substances from inside of the cell. In order to gain insight into how the cells are responding to the

environment, expression levels due to treatment effects of each gene encoding a binding protein were plotted as a function of time after heat shock for *P. furious* (Figure 10-11) and *T. maritima* (Figure 12). While realizing that annotations present in the genome sequence data may be inaccurate, it is apparent that expression of genes encoding ABC transporter binding proteins over time depends strongly on growth substrate and organism. Most transporters involving sugar or di/oligopeptide uptake decreased significantly over time after heat shock when *P. furiosus* was grown on maltose, but expression of sugar and di/oligopeptide binding proteins was relatively steady when *P. furiosus* was grown proteolytically. Genes encoding binding proteins for ABC transport systems predicted to transport sugars or di/oligopeptides in *T. maritima* remained steady or (in most cases) increased as heat shock was applied. This difference in strategy may be due to the relative importance of compatible solutes at the two different temperatures or how compatible solutes are obtained by each organism. With one exception, all transporters predicted to transport inorganic solutes showed relatively unchanged levels of expression over time. In *P. furiosus*, a Mn^{2+}/Zn^{2+} transporter showed decreased expression over the time course of heat shock.

Overall, as diagrammed in Figure 13, heat shock response by *P. furiosus* involved similar numbers of genes in functional categories over the first hour under both tested growth environments. A wide divergence existed at 90 min. in P. furiosus; for growth on tryptone, gene expression across all functional categories increased at 90 min. while gene expression across functional categories decreased at 90 min. for growth on maltose (see Figure 13). Early in heat shock, the transcription of genes involved in DNA replication, recombination and repair were more rapidly elevated in *P. furiosus* than *T. maritima*. Also, unlike *P. furiosus*, *T. maritima* appears to down-regulate the expression of genes involved in cell

motility and secretion early in the heat shock response. The majority of genes involved in proteolytic fermentation, the modified Embden-Myerhoff and Entner-Douderoff pathways, gluconeogenesis and the pentose phosphate pathway were downregulated throughout the course of the experiment in both organisms for both growth media (data not shown).

Conclusions

Though *P. furiosus* and *T. maritima* originate from different domains of life, both organisms show a time-dependent response to thermal stress. This dynamic information on the induction and development of the heat shock response presented here was able to elucidate differences between immediate and long-term survival strategies in this organism. In addition, cluster analyses reveal that while certain genes in these organisms were differentially expressed to various extents upon heat shock, maintenance of pre-heat shock expression levels can be equally significant. Heat shock in both of these hyperthermophiles proceeds through the immediate up-regulation of genes encoding important molecular chaperones and proteins involved in the stabilization and repair of DNA as well as a latent induction of genes that may be involved in a final effort to stabilize metabolic integrity. Genes involved in energy metabolism are consistently down-regulated throughout the time course of thermal stress. This also raises questions concerning the set of proteases mediating protein turnover under thermal stress and how the intracellular concentration of misfolded and abnormal proteins in hyperthermophiles and less thermophilic organisms compares. In contrast to mesophilic bacteria, both organisms do not immediately induce the expression of genes encoding ATP-dependent proteases. Though media formulation can influence the heat shock response in *P. furiosus*, core heat shock genes are differentially regulated on both

media and the results obtained here compared well with those reported by Shockley et al. who used a targeted microarray to study the effect of heat shock in *P. furiosus* after one hour (26). A central issue here is the extent to which thermal stress perturbs cellular function in hyperthermophiles.

Acknowledgements

This work was supported in part by grants from the Department of Energy (Energy Biosciences Program) and the National Science Foundation (Biotechnology Program). KRS acknowledges support from a Department of Education GAANN Fellowship. SBC acknowledges support from a NIEHS Traineeship in Bioinformatics.

References

1. **Bell, S. D., and S. P. Jackson.** 1998. Transcription and translation in Archaea: a mosaic of eukaryal and bacterial features. Trends Microbiol. **6:**222-228.
2. **Chhabra, S. R., K. R. Shockley, S. B. Conners, K. L. Scott, R. D. Wolfinger, and R. M. Kelly.** 2003. Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima*. J. Biol. Chem. **278:**7540-7552.
3. **Duguet, M.** 1993. The helical repeat of DNA at high temperature. Nucleic Acids Res. **21:**463-468.
4. **Grogan, D. W.** 1998. Hyperthermophiles and the problem of DNA instability. Mol. Microbiol. **28:**1043-1049.
5. **Gupta, R. S.** 1998. What are archaebacteria: life's third domain or monoderm prokaryotes related to gram-positive bacteria? A new proposal for the classification of prokaryotic organisms. Mol. Microbiol. **29:**695-707.
6. **Hasseman, J.** 2001, posting date. TIGR Microarray Protocols. [Online.]
7. **Helmann, J. D., M. F. Wu, P. A. Kobel, F. J. Gamo, M. Wilson, M. M. Morshedi, M. Navre, and C. Paddon.** 2001. Global transcriptional response of *Bacillus subtilis* to heat shock. J. Bacteriol. **183:**7318-28.
8. **Hengge-Aronis, R.** 1996. Back to log phase: sigma S as a global regulator in the osmotic control of gene expression in *Escherichia coli*. Mol. Microbiol. **21:**887-893.
9. **Hobbie, J. E., R. J. Daley, and S. Jasper.** 1977. Use of nuclepore filters for counting bacteria by fluorescence microscopy. Appl. Environ. Microbiol. **33:**1225-1228.

10. **Jolivet, E., F. Matsunaga, Y. Ishino, P. Forterre, D. Prieur, and H. Myllykallio.** 2003. Physiological responses of the hyperthermophilic archaeon "*Pyrococcus abyssi*" to DNA damage caused by ionizing radiation. J. Bacteriol. **185:**3958-3961.

11. **Komori, K., T. Miyata, H. Daiyasu, H. Toh, H. Shinagawa, and Y. Ishino.** 2000. Domain analysis of an archaeal RadA protein for the strand exchange activity. J. Biol. Chem. **275:**33791-33797.

12. **Koning, S. M., W. N. Konings, and A. J. M. Driessen.** 2002. Biochemical evidence for the presence of two a-glucosidase ABC-transport systems in the hyperthermophilic archaeon *Pyrococcus furiosus*. Archaea **1:**19-25.

13. **Koonin, E. V., L. Aravind, and M. Y. Galperin.** 2000. A comparative-genomic view of the microbial stress response. ASM Press, Washington, D. C.

14. **Lamosa, P., L. O. Martins, M. S. Da Costa, and H. Santos.** 1998. Effects of temperature, salinity, and medium composition on compatible solute accumulation by *Thermococcus* spp. Appl. Environ. Microbiol. **64:**3591-3598.

15. **Leigh, J. A.** 1999. Transcriptional regulation in Archaea. Curr. Opin. Microbiol. **2:**131-134.

16. **Lopez Garcia, P., and P. Forterre.** 2000. Thermal stress in hyperthermophiles. ASM Press, Washington, D. C.

17. **Macario, A. J. L., and E. C. de Macario.** 2001. The molecular chaperone system and other anti-stress mechanisms in archaea. Front. Biosci. **6:**d262-263.

18. **Petersohn, A., M. Brigulla, S. Haas, J. D. Hoheisel, U. Volker, and M. Hecker.** 2001. Global analysis of the general stress response of *Bacillus subtilis*. J. Bacteriol. **183:**5617-5631.

19. **Pysz, M. A., D. E. Ward, K. R. Shockley, C. I. Montero, S. B. Conners, M. R. Johnson, and R. M. Kelly.** 2004. Transcriptional analysis of dynamic heat-shock response by the hyperthermophilic bacterium *Thermotoga maritima*. Extremophiles.

20. **Richmond, C. S., J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner.** 1999. Genome-wide expression profiling in *Escherichia coli* K-12. Nucleic Acids Res. **27:**3821-3835.

21. **Sambrook, J., E. H. Fritsh, and T. Maniatis.** 1989. Extraction, purification, and analysis of messenger RNA from eukarytoic cells. Cold Spring Harbor Laboratory Press, Plainview, NY.

22. **Santos, H., and M. S. da Costa.** 2002. Compatible solutes of organisms that live in hot saline environments. Environ. Microbiol. **4:**501-509.

23. **Schumann, W.** 2003. The *Bacillus subtilis* heat shock stimulon. Cell Stress Chaperones **8:**207-217.

24. **Schut, G. J., S. D. Brehm, S. Datta, and M. W. Adams.** 2003. Whole-genome DNA microarray analysis of a hyperthermophile and an archaeon: *Pyrococcus furiosus* grown on carbohydrates or peptides. J. Bacteriol. **185:**3935-3947.

25. **Schut, G. J., J. Zhou, and M. W. Adams.** 2001. DNA microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus*: evidence for anNew type of sulfur-reducing enzyme complex. J. Bacteriol. **183:**7027-7036.

26. **Shockley, K. R., D. E. Ward, S. R. Chhabra, S. B. Conners, C. I. Montero, and R. M. Kelly.** 2003. Heat shock response by the hyperthermophilic archaeon *Pyrococcus furiosus*. Appl. Environ. Microbiol. **69:**2365-2371.

27. **Soppa, J.** 1999. Transcription initiation in Archaea: facts, factors and future aspects. Mol. Microbiol. **31:**1295-1305.

28. **Stintzi, A.** 2003. Gene expression profile of *Campylobacter jejuni* in response to growth temperature variation. J. Bacteriol. **185:**2009-2016.

29. **Struhl, K.** 1999. Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. Cell **98:**1-4.

30. **Tatusov, R. L., E. V. Koonin, and D. J. Lipman.** 1997. A genomic perspective on protein families. Science **278:**631-637.

31. **Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin.** 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29:**22-28.

32. **Thomm, M.** 1996. Archaeal transcription factors and their role in transcription initiation. FEMS Microbiol. Rev. **18:**159-171.

33. **Wolfinger, R. D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules.** 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. **8:**625-637.

34. **Yura, T., M. Kanemoi, and T. Morita.** 2000. Bacterial Stress Responses. ASM Press, Washington, D. C.

35. **Yura, T., and K. Nakahigashi.** 1999. Regulation of the heat-shock response. Curr. Opin. Microbiol. **2:**153-158.

Tables

Table 1. Dynamic Heat Shock Results

	Up-regulated					Down-regulated				
Time (min.)	**0**	**5**	**30**	**60**	**90**	**0**	**5**	**30**	**60**	**90**
P. furiosus										
Tryptone	**249**	**154**	**246**	**137**	**405**	**281**	**161**	**285**	**216**	**560**
Maltose	**275**	**177**	**143**	**87**	**51**	**368**	**113**	**179**	**353**	**428**
Shared	**108**	**77**	**71**	**30**	**38**	**91**	**34**	**103**	**111**	**245**
T. maritima										
Maltose	**150**	**213**	**230**	**272**	**352**	**175**	**259**	**223**	**290**	**430**

P<0.001

***P. furiosus*; 2016 total ORFs on array (>99%)**
***T. maritima*; 1908 total ORFs on array (>99%)**

Table 2. Consistently Up-regulated Genes

Description	Locus
P. furiosus **(60 minutes)**	
hypothetical proteins	**PF0624*, PF0952*, PF1372**
chaperones	**PF1883 [small heat shock protein]**
Aminotransferase	**PF1906 [adenosylmethionine-8-amino-7-oxononanoate aminotransferase]**
T. maritima **(90 minutes)**	
hypothetical proteins	**TM0002, TM0003, TM0004, TM0063, TM0069, TM0070, TM0180, TM0375, TM0390, TM0582, TM0660, TM0706, TM0979, TM0980, TM0981, TM0982, TM0983, TM0989, TM1145, TM1412, TM1420, TM1455, TM1593, TM1659**
ribosomal proteins	**TM1454 [50S ribosomal protein L13]**
	TM1505 [30S ribosomal protein S12]
	TM1591 [50S ribosomal protein L20]
	TM1592 [50S ribosomal protein L35]
chaperones	**TM0373 [dnaK protein]**
	TM0374 [heat shock protein, class I]
	TM0505 [groES protein]
	TM0506 [groEL protein]
	TM0849 [dnaJ protein]
transport	**TM0123 [zinc transport system substrate-binding protein]**
	TM0125 [zinc transport system permease protein]
	TM0300 [putative oligopeptide transport system substrate-binding protein]
	TM0389 [putative ABC-2 type transport system ATP-binding protein]
	TM1232 [multiple sugar transport system ATP-binding protein]
	TM1276 [putative multiple sugar transport system ATP-binding protein]
DNA repair	**TM0480 [excinuclease ABC, subunit A]**
	TM1546 [single stranded DNA-specific exonuclease]
regulators	**TM0510 [iron-dependent transcriptional repressor, putative]**
	TM1776 [ferric uptake regulation protein]

Table 2 (cont.)

other	**TM0024 [laminarinase]**
	TM0064 [uronate isomerase, putative]
	TM0097 [nicotinate-nucleotide adenylyltransferase]
	TM0219 [flagellar export/assembly protein]
	TM0873 [frameshift]
	TM1267 [thiH protein, putative]
	TM1418 [frameshift]
	TM1576 [hemolysin]
	TM1598 [RNA polymerase sigma-E factor]
	TM1658 [S-adenosylmethionine synthetase]
	TM1705 [lysyl-tRNA synthetase]
	16SrRNA [rRNA]

***ORFs in P. furiosus which were up-regulated for the full 90 minutes**

Table 3. Consistently Down-regulated Genes

Description	Locus
P. furiosus **(60 minutes)**	
hypothetical proteins	**PF0152, PF1786**
	PF0043 [phosphoenolpyruvate synthase]
	PF1800 [adenylate kinase (ATP-AMP transphosphorylase)]
T. maritima **(90 minutes)**	
hypothetical proteins	**TM0035, TM0329, TM0619, TM0644, TM0693, TM0696, TM0866, TM0992, TM1020, TM1024, TM1083, TM1252, TM1266, TM1337, TM1682, TM1690, TM1807, TM1810, TM1813, TM1872**
ATP synthesis	**TM1611 [ATP synthase F1, subunit gamma]**
	TM1612 [ATP synthase F1, subunit alpha]
	TM1613 [ATP synthase F1, subunit delta]
regulators	**TM0275 [transcriptional regulator, GntR family]**
	TM1081 [anti-sigma factor antagonist, putative]
	TM1082 [lexA repressor]
transport and metabolism	**TM0460 [putative oligopeptide transport system s]**
	TM0501 [putative oligopeptide transport system ATP-binding protein]
	TM0549 [acetolactate synthase small subunit]
	TM1400 [putative aminotransferase]
	TM1518 [aspartokinase II]
	TM1519 [2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase]
	TM1520 [dihydrodipicolinate reductase]
	TM1522 [diaminopimelate epimerase]
	TM0404 [deoxycytidylate deaminase, putative]
	TM1245 [phosphoribosylformylglycinamidine synthase I]
	TM1246 [phosphoribosylformylglycinamidine synthase II]
	TM1248 [phosphoribosylglycinamide formyltransferase]
	TM1249 [phosphoribosylaminoimidazolecarboxamide formyltransferase / I]
	TM1250 [phosphoribosylamine--glycine ligase]
	TM1251 [phosphoribosylformylglycinamidine cyclo-ligase]

Table 3 (cont.)

energy production and conservation	**TM0274 [acetate kinase]**
	TM1426 [NADH dehydrogenase I chain G]
other	**TM0273 [fructose-bisphosphate aldolase]**
	TM0328 [m4C-methyltransferase]
	TM0521 [heat shock protein HslV]
	TM0630 [nucleotide sugar epimerase, putative]
	TM0718 [purine-binding chemotaxis protein]
	TM0797 [probable 2-phosphosulfolactate phosphatase]
	TM0824 [astB/chuR-related protein]
	TM1253 [NH(3)-dependent NAD(+) synthetase]
	TM1254 [beta-phosphoglucomutase, putative]
	TM1589 [clostripain-related protein]
	TM1765 [N utilization substance protein B]
	TM1803 [dnaJ-related protein]
	TM1870 [septum site-determining protein MinD]

Table 4. Heat Shock Response in *P. furiosus

Locus	Description			Time (min.) Tryptone + S^0					Maltose + yeast extract				
		60 min.											
	P. furiosus	HS_1	HS_2	0	5	30	60	90	0	5	30	60	90
chaperones													
PF1883	Small heat shock protein	>7.7	>6.9	+	+	+	+	+	+	+	+	+	
PF1974	Thermosome	>4.0	>4.2	+	+	+		-	+	+		+	+
PF1882	VAT	6.9	3.7	+	+	+		+	+	+	+		
PF0963	VAT	4.6	2.7	+	+	+			+				
ATP-dependent proteases													
PF1404	Psm β-1	2.0	2.0	+									
PF0159	Psm β-2	1.9	1.4				-	-					
PF0115	PAN	1.2	-1.1					-					
PF1571	Psm α	-3.9	-2.2	-		-							
PF0467	Lon	-2.9	>-1.3					-			-		
ATP-independent proteases													
PF0688	Subtilisin-like protease	7.0	16.5			+	+	+			+	+	
PF1185	ArgE/peptidase	4.5	3.7	+				-	+	+	+		
PF1597	HtpX heat shock protein	4.5	3.0						-	-			
PF0287	Pyrolysin	>-4.9	-2.9	-			-	-	-				
Carbohydrate Processing and Utilization													
PF0073	CelB	18.1	9.5			+	+	+	+	+	+	+	
PF0137	Putative methyltransferase (deacetylase)	11.1	3.6			+		+			+		
PF0076	LamA	6.4	3.8			+				+	+		
PF1234	ChitB	4.4	6.2	+		+	+	+	+	+	+	+	

Table 4 (cont.)

Compatible Solutes													
PF1739	Trehalose/maltose binding protein	6.5	2.4	-				-	-	-	-	-	
PF1742	Putative trehalose synthase	2.7	4.0	+					-	-			
PF0127	Spermidine synthase	2.7	4.0										
PF1616	Myo-inositol-1,1-phosphate synthase	N/A	N/A	+		+			+	+			
DNA Repair													
PF1926	RadA	>2.6	2.0	+	+		-	+	+				
PF0021	RadB	1.4	-1.2										
PF2019	RP-A subunit homolog	N/A	N/A	+	+	+		+	+				
PF1578	Type II DNA topoisomerase subunit a	N/A	N/A	+	+			+	+		-		
PF1579	DNA topoisomerase VI, subunit b	N/A	N/A	+	+		-	+	+				
PF0126	DNA repair protein rad25	N/A	N/A	+		+	-	+	+		+	-	
PF2015	ATP-dependent RNA helicase, putative	N/A	N/A	+	+			+		+			

*P<0.01

HS_1 and HS_2 refer to replicates from a previous study (26) in which *P. furiosus* was grown on tryptone and S^0 and yeast extract and differential gene expression was monitored after an hour of thermal stress.

Table 5. Known and Putative Chaperones

Locus	Description	Time (minutes)									
		Tryptone + S^0					Maltose + yeast extract				
		0	5	30	60	90	0	5	30	60	90
	P. furiosus										
PF1883	Small heat shock protein	+	+	+	+	+	+	+	+	+	
PF1974	Thermosome	+	+	+		-	+	+			
PF0375	Prefoldin (subunit α)	+	+	+	-	-					
PF0382	Prefoldin (subunit β)	-			-	-					
PF1882	VAT	+	+	+			+	+			
PF0963	VAT	+	+	+			+				
PF0467	Lon										
PF0741	Thioredoxin										
PF0094	Glutaredoxin				-	-	-			-	-
PF1401	PPIases	+				-				-	-
PF1030	CsaA					-					
PF1667	CsaA										
PF1167	Archaeal Hop? (smc-like)										
PF1879	Archaeal Hop?	-									
PF1966	Archaeal Hop?	-									
PF1965	Archaeal Hop?				-	-	-				
PF0060	Achaeal Hip?				-					-	
PF0341	Archaeal Hip?				+	+					
PF0335	Archaeal Hip?						+				
PF0172	Archaeal Hip?										
PF1932	Archaeal BAG-1?	+					+				
PF1545	Archaeal NAC?										

Table 5 (cont.)

	T. maritima					
TM0373	DnaK	+	+	+	+	+
TM0374	Small heat shock protein	+	+	+	+	+
TM0505	GroES	+	+	+	+	+
TM0506	GroEL	+	+	+	+	+
TM0849	DnaJ	+	+	+	+	+
TM0850	GrpE	+	+	+		+

*P<0.01

Italicized ORFs in *P. furiosus* were identified in (17) as possible eukaryal-like chaperones present in the archaea.

Figures

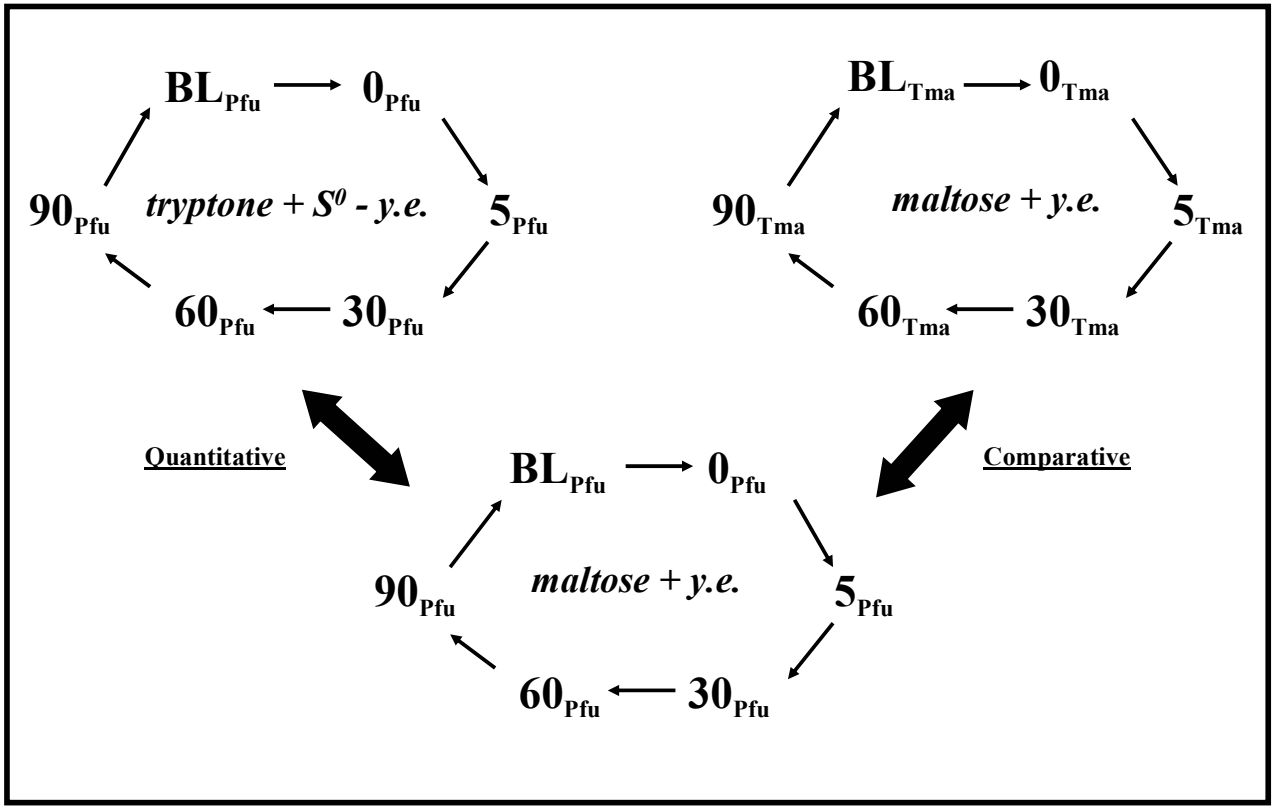


Figure 1. Loop design for the study of dynamic heat shock response. The arrow ends correspond to the Cy3 and Cy5 channels as follows: Cy3 •→Cy5.

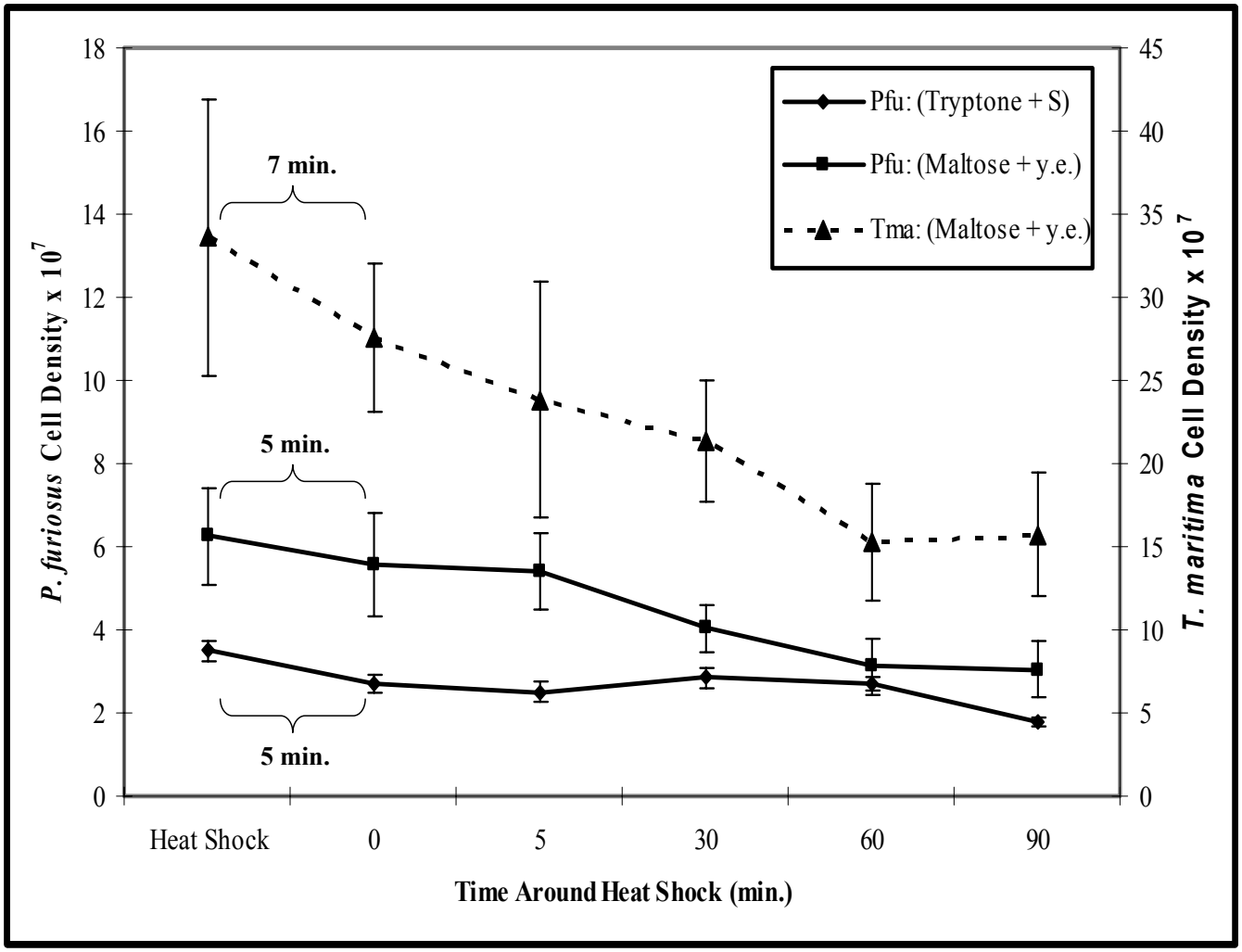


Figure 2. Response of both cells growing until mid-exponential growth phase to a temperature shock. *Pyrococcus furiosus* growing at 90°C was subjected to a temperature of 105°C. *Thermotoga maritima* growing at 80°C and was subjected to a temperature of 90°C.

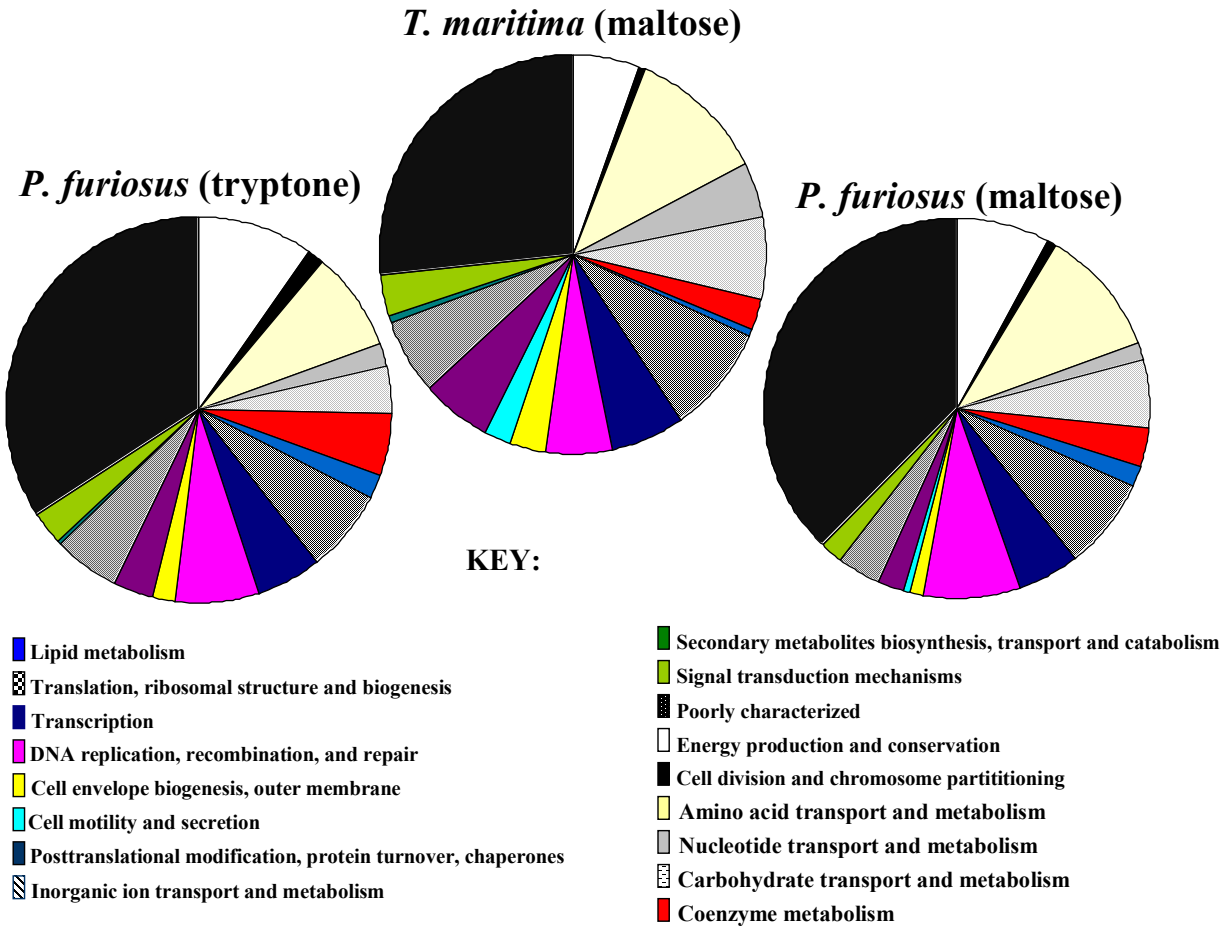


Figure 3. Distribution of the functional categories for differentially expressed ORFs. Shown is induction of genes by functional category (30, 31) after 5 minutes of thermal stress.

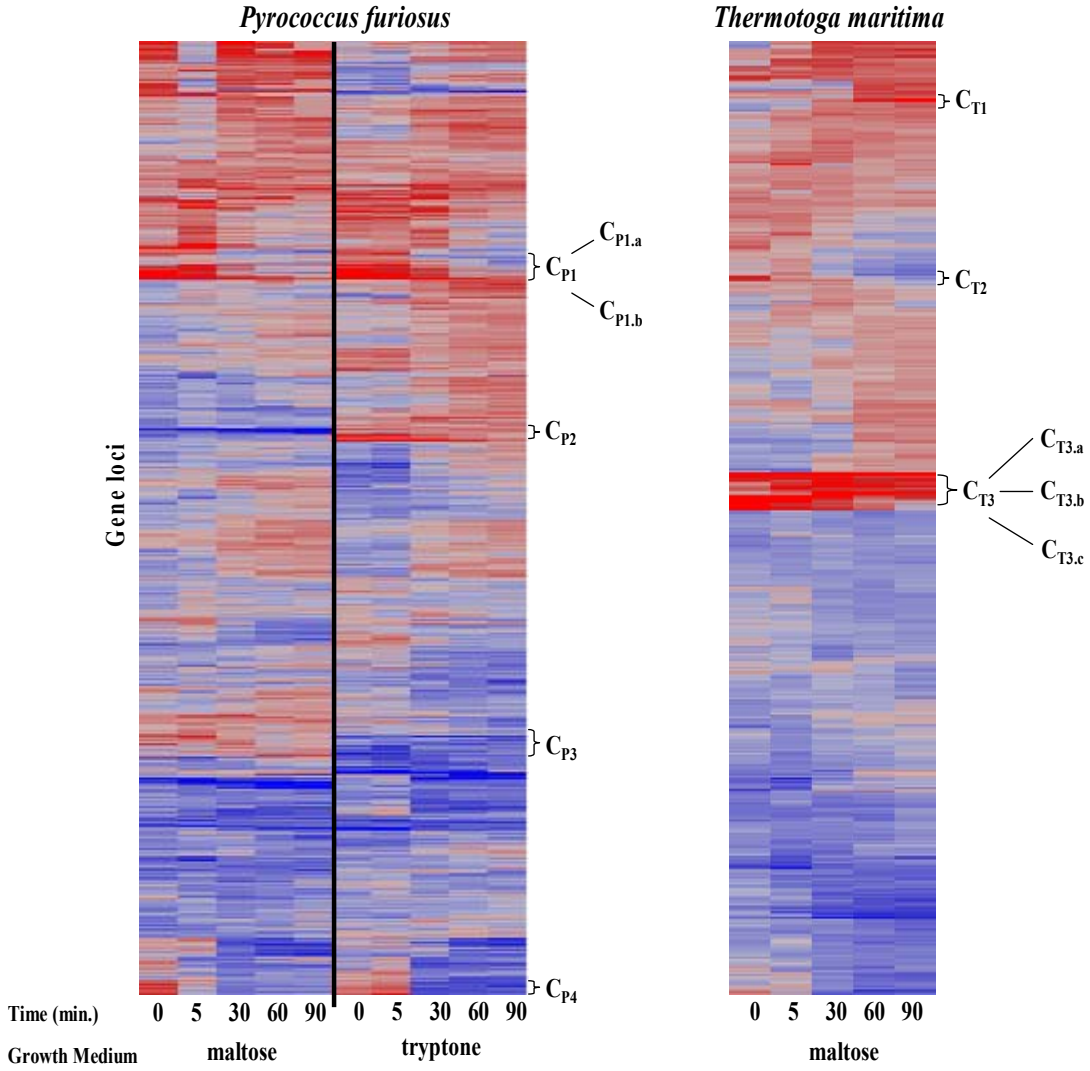


Figure 4. Hierarchical clustering patterns based on fold change showing the most dramatic ORFs induced immediately upon heat shock. Fold change is based on differences in least squares mean estimates of treatment effects, which is analogous to $\log_2$ transformed fold changes. Rows in the clusters represent genes and columns represent time points since the application of heat shock. Clusters C_{P1}-C_{P4} and C_{T1}-C_{T3} are shown in greater detail in Figures 5-7.

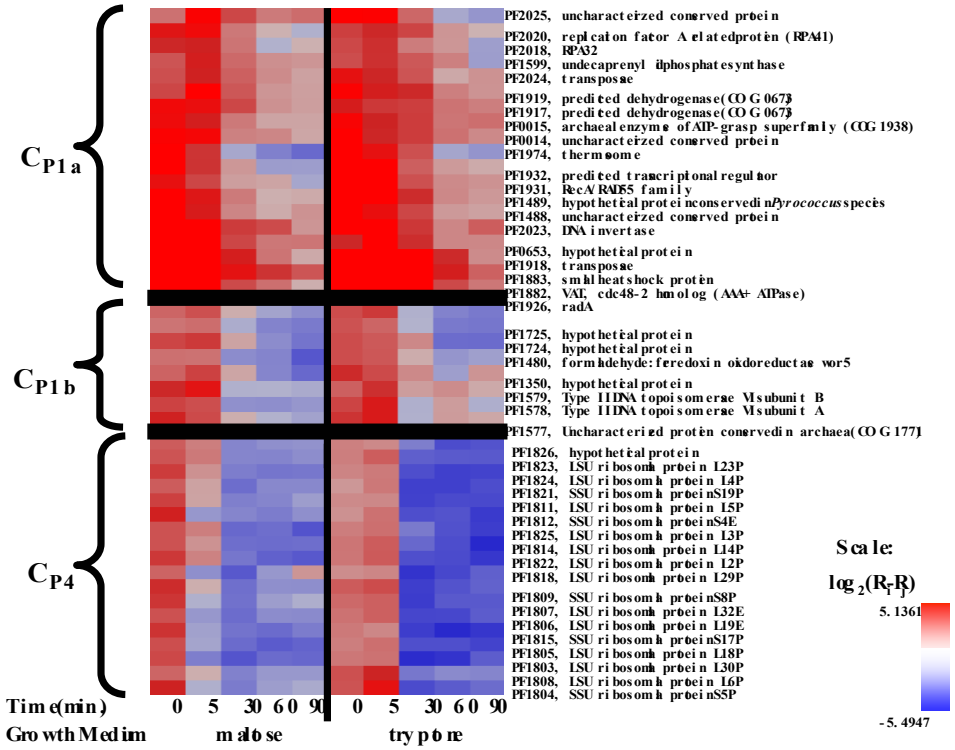


Figure 5. Immediate induction of gene expression in *P. furious*. Fold change is based on differences in least squares means estimates. Known or putative functions as they appear in the genome sequence are indicated.

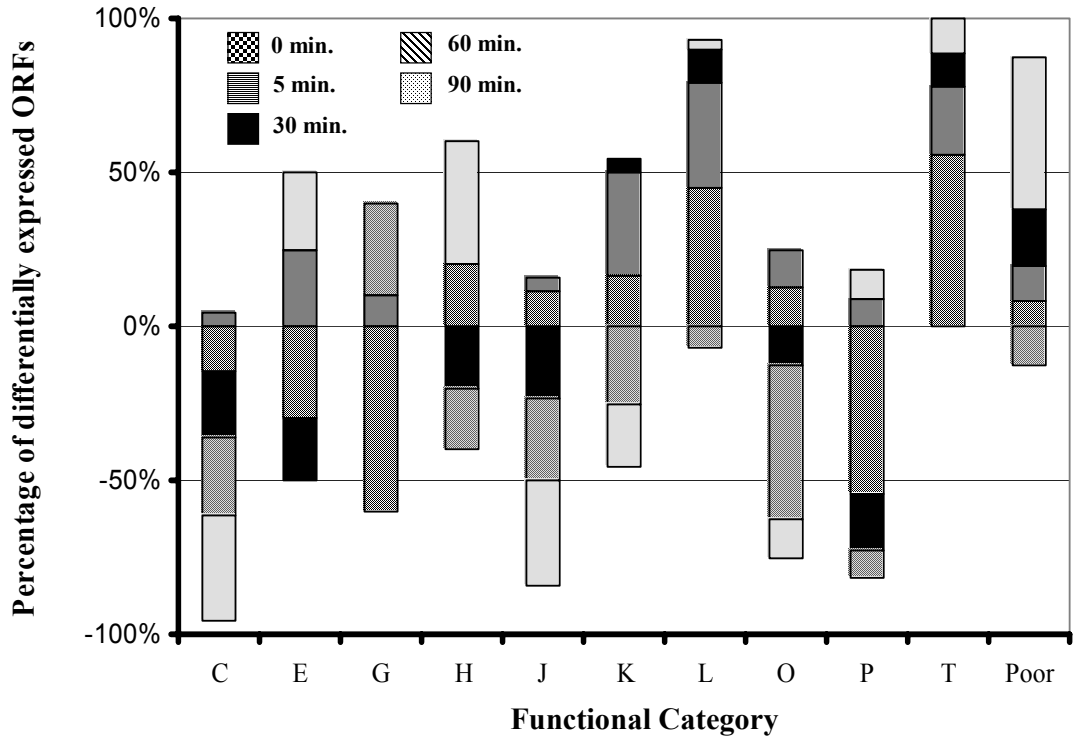


Figure 6. Percent of ORFs differentially regulated at each time point in *P. furiosus* by NCBI functional category. Shown are differences in the numbers of differentially expressed ORFs (P<0.001) in *P. furiosus* for growth on maltose and tryptone as a function of time after heat shock was applied (0, 5, 30, 60, or 90 minutes). Percentages in the positive direction indicate that more ORFs that were up-regulated than down-regulated; percentages in the negative direction indicate the numbers of ORFs that were down-regulated over the number of ORFs that were up-regulated. Functional categories are derived from the framework of clusters of orthogolous genes available from NCBI (30, 31).

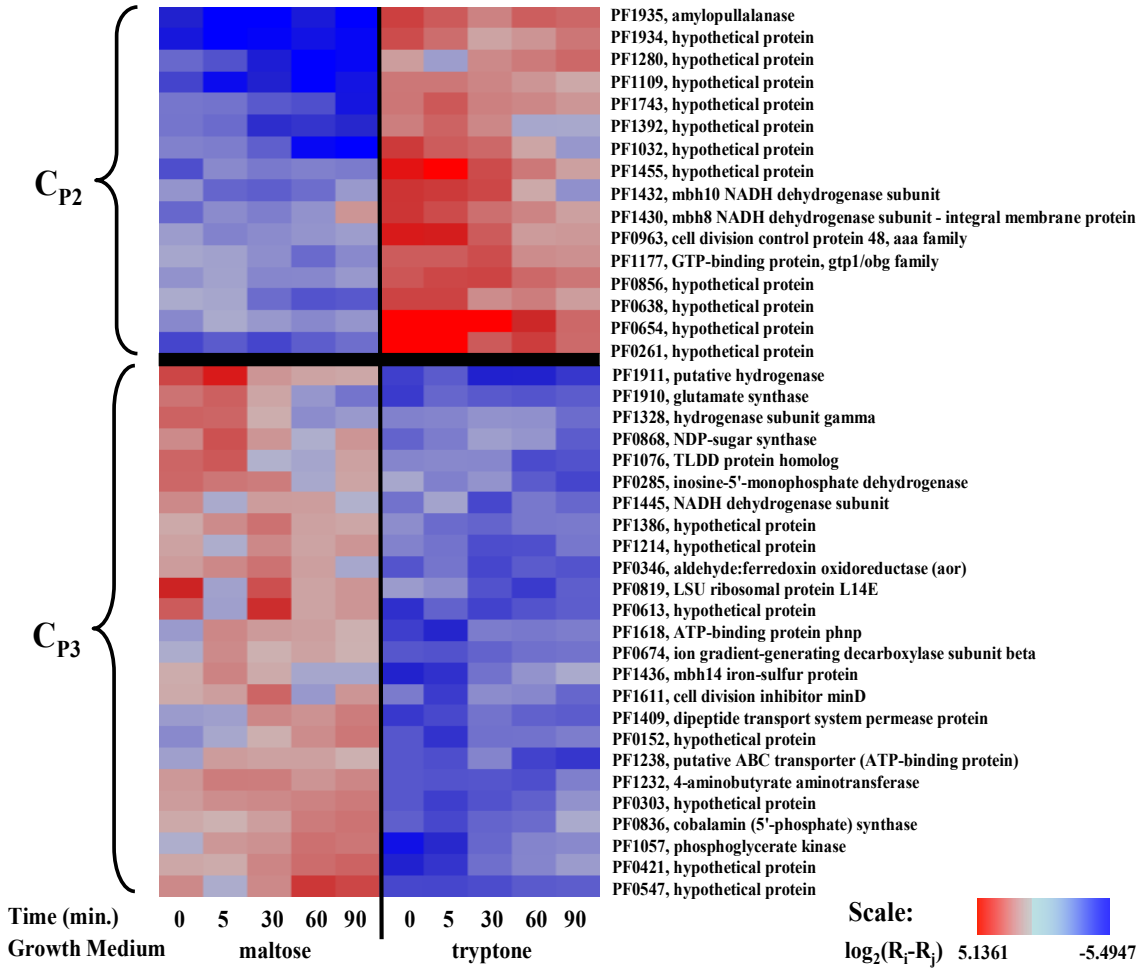


Figure 7. Medium-dependent induction of gene expression in *P. furiosus*. Shown are differences in least squares means estimates of gene-specific treatment effects. Known or putative functions as they appear in the genome sequence are indicated.

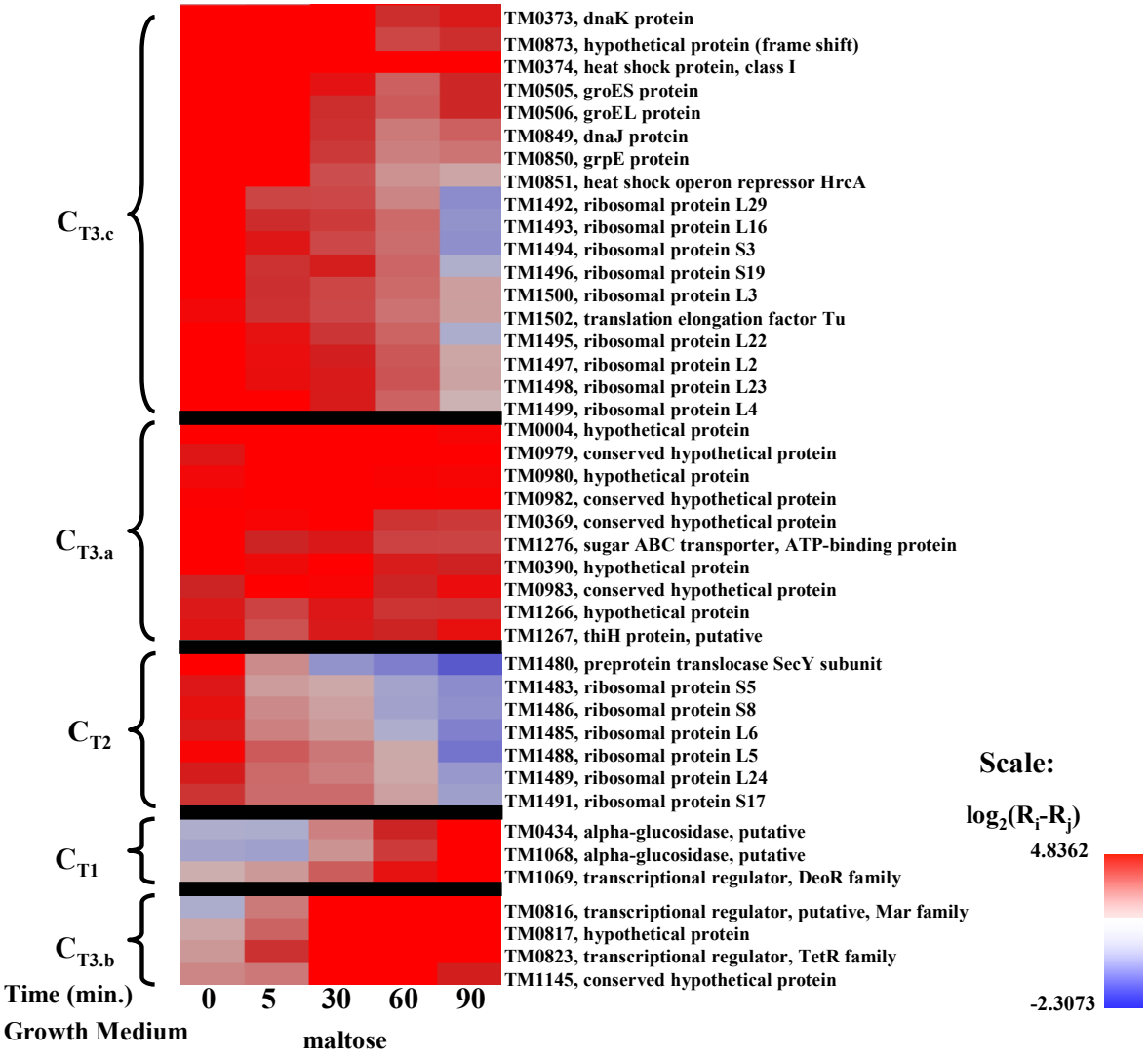


Figure 8. Immediate and long-term gene expression in *T. maritima*. Shown are differences in least squares means estimates. Known or putative functions as they appear in the genome sequence are indicated.

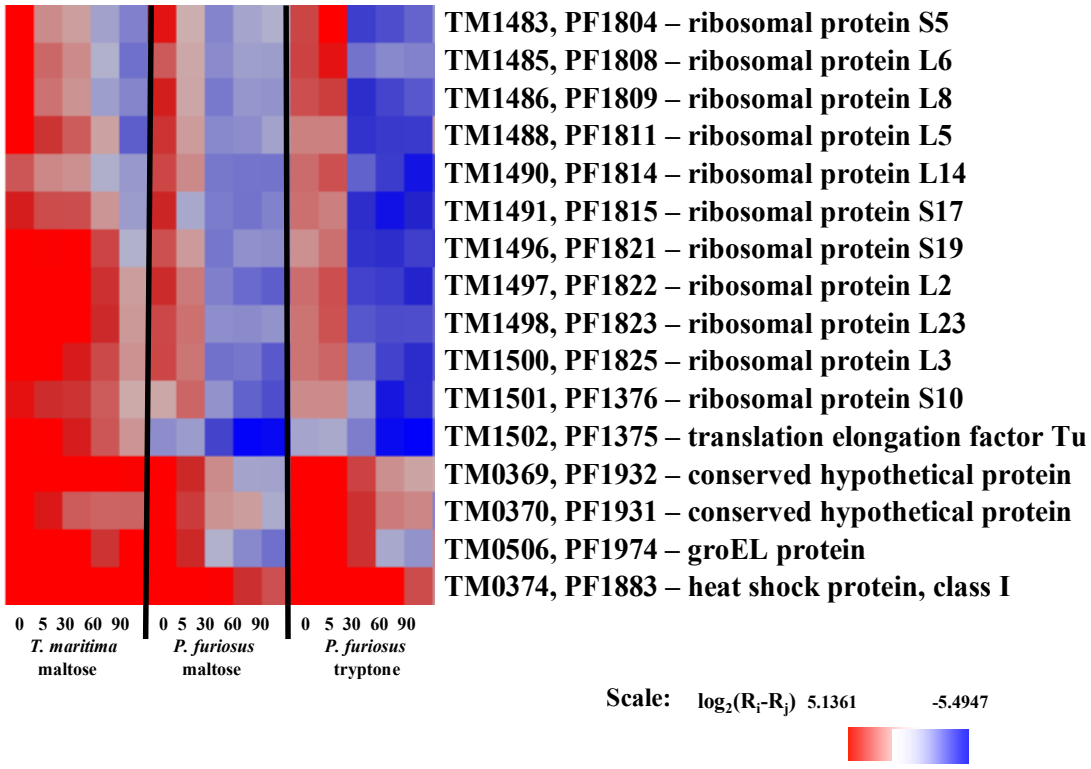


Figure 9. Immediate induction of homologous open reading frames in *P. furiosus* and *T. maritima*. Shown are paired sequence homologs with >40% sequence similarity between *P. furiosus* and *T. maritima* that were upregulated in response to heat shock.

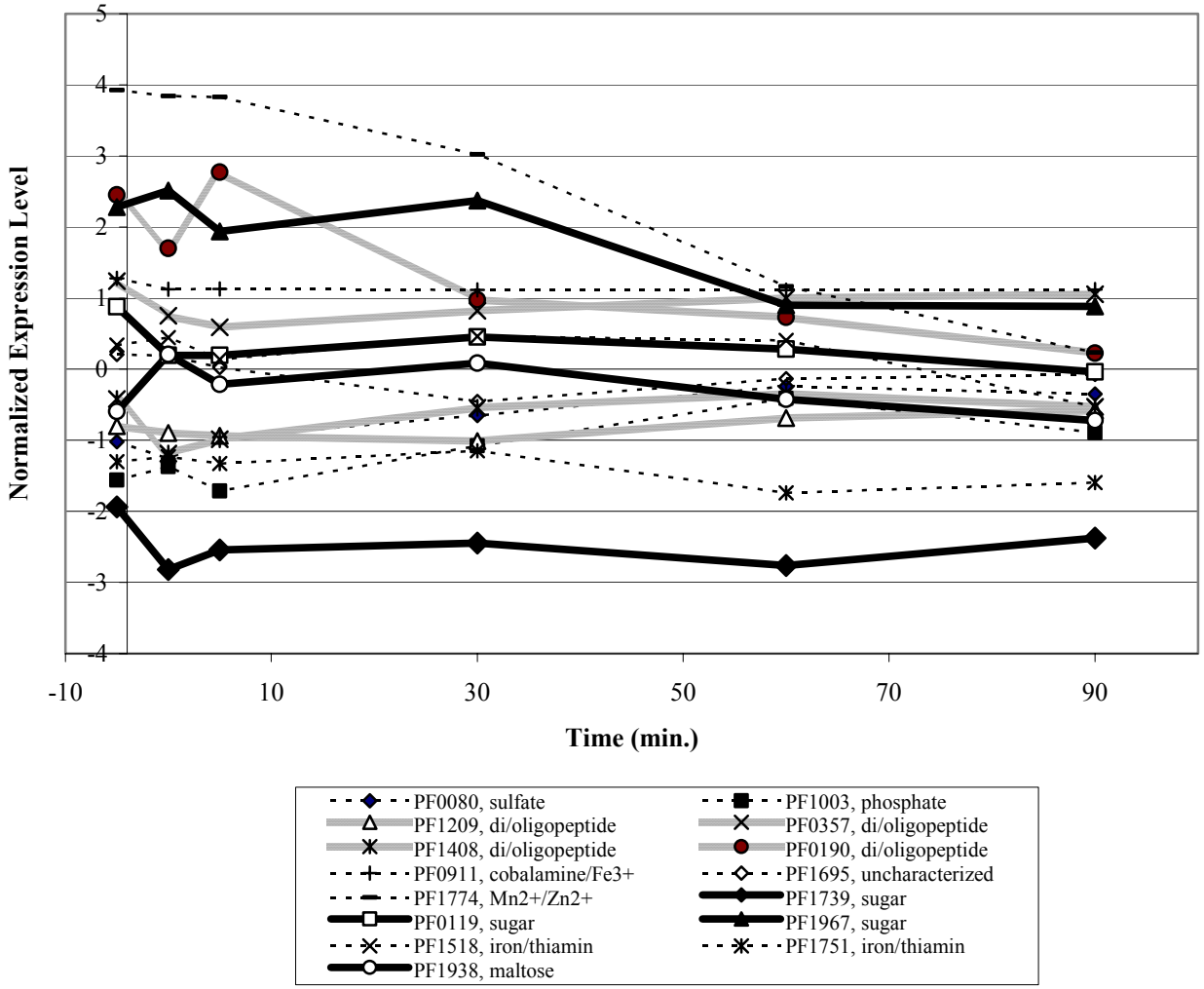


Figure 10. Least square means estimates of binding proteins in *P. furiosus* grown on tryptone. Dynamic trends in least squares means estimates of expression level due to treatments for binding proteins in ABC transport gene clusters.

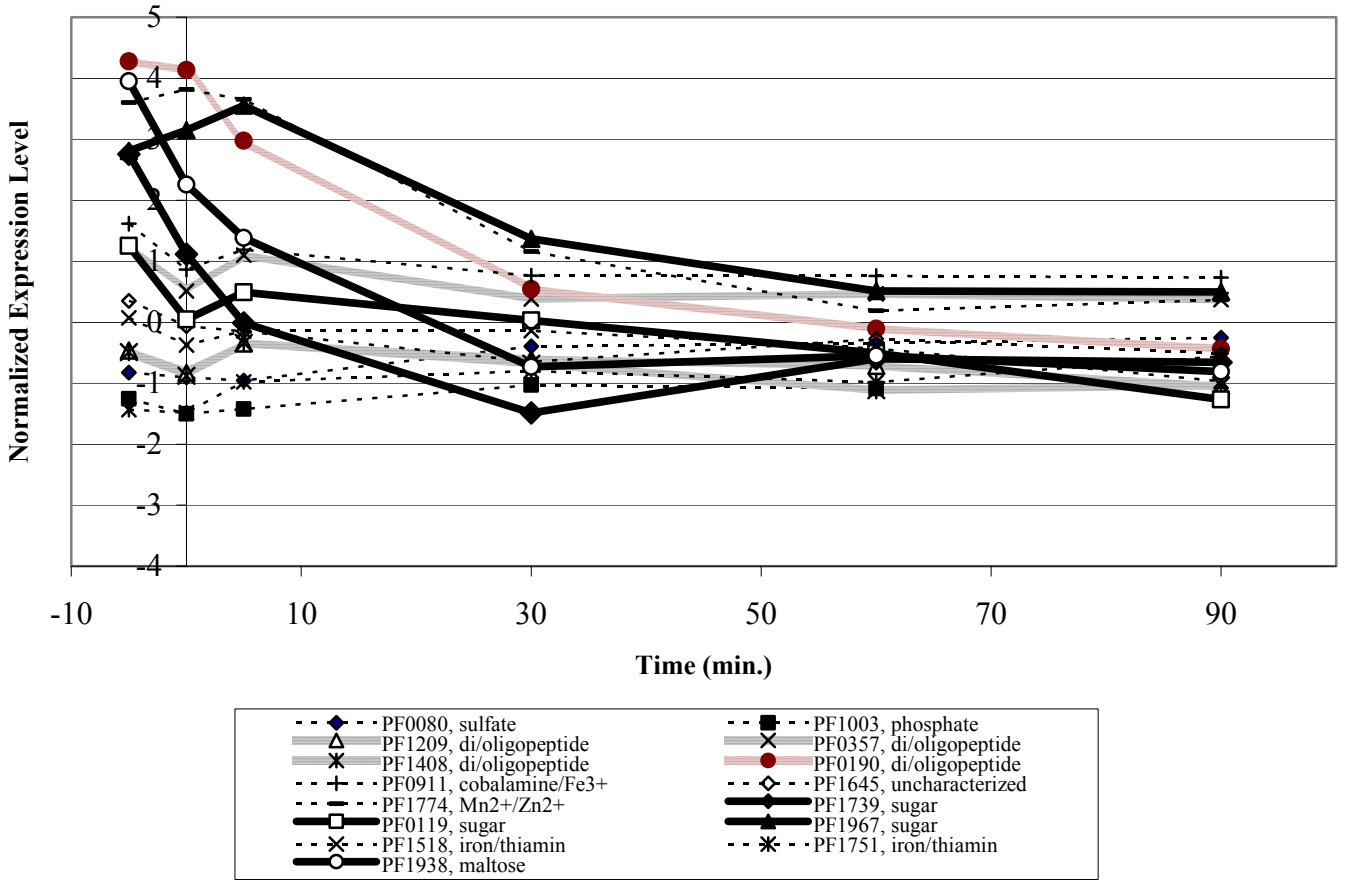


Figure 11. Least square means estimates of binding proteins in *P. furiosus* grown on maltose. Dynamic trends in least squares means estimates of expression level due to treatments for binding proteins in ABC transport gene clusters.

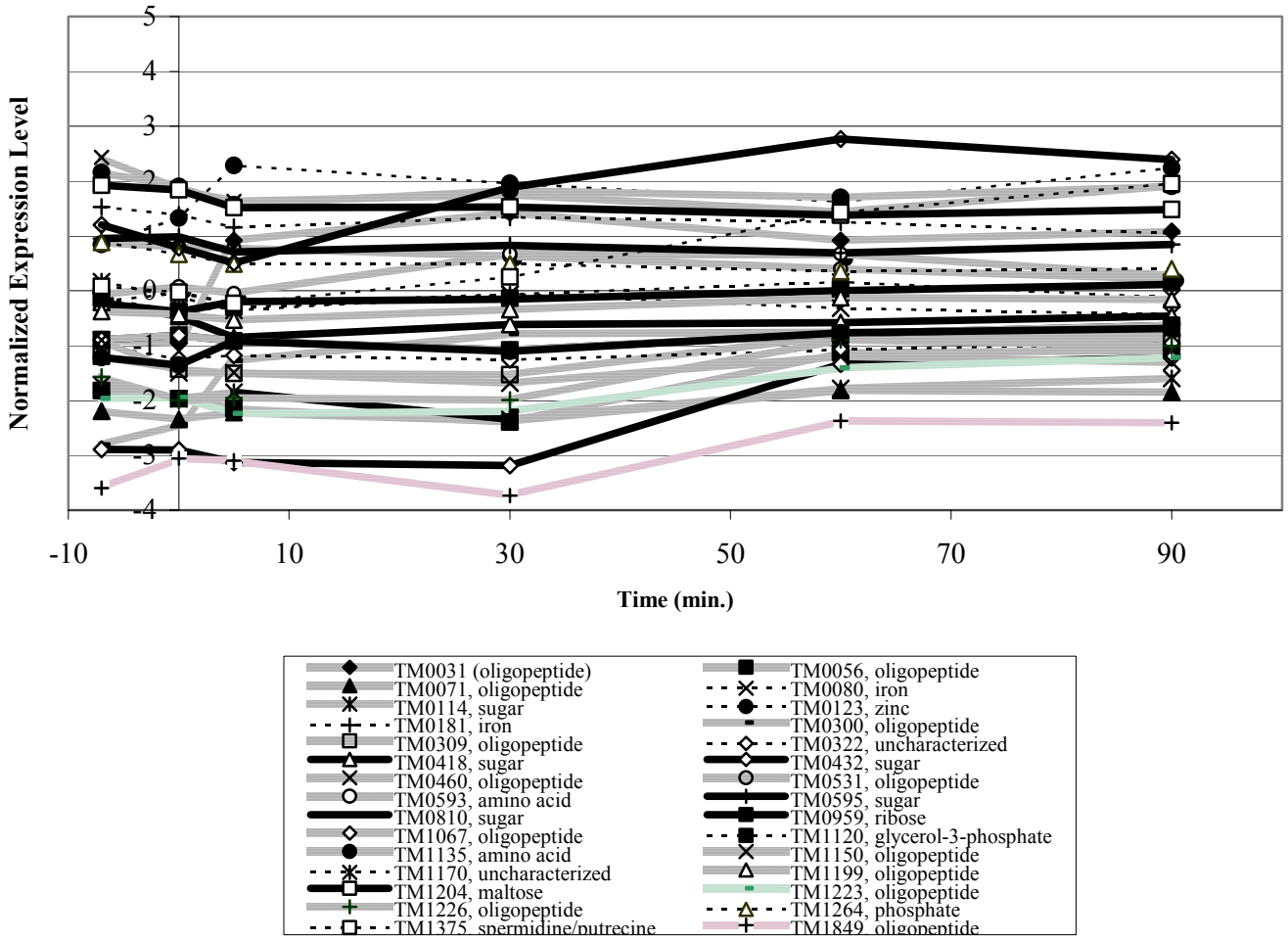


Figure 12. Least square means estimates of binding proteins in *T. maritima* grown on maltose. Dynamic trends in least squares means estimates of expression level due to treatments for binding proteins in ABC transport gene clusters.

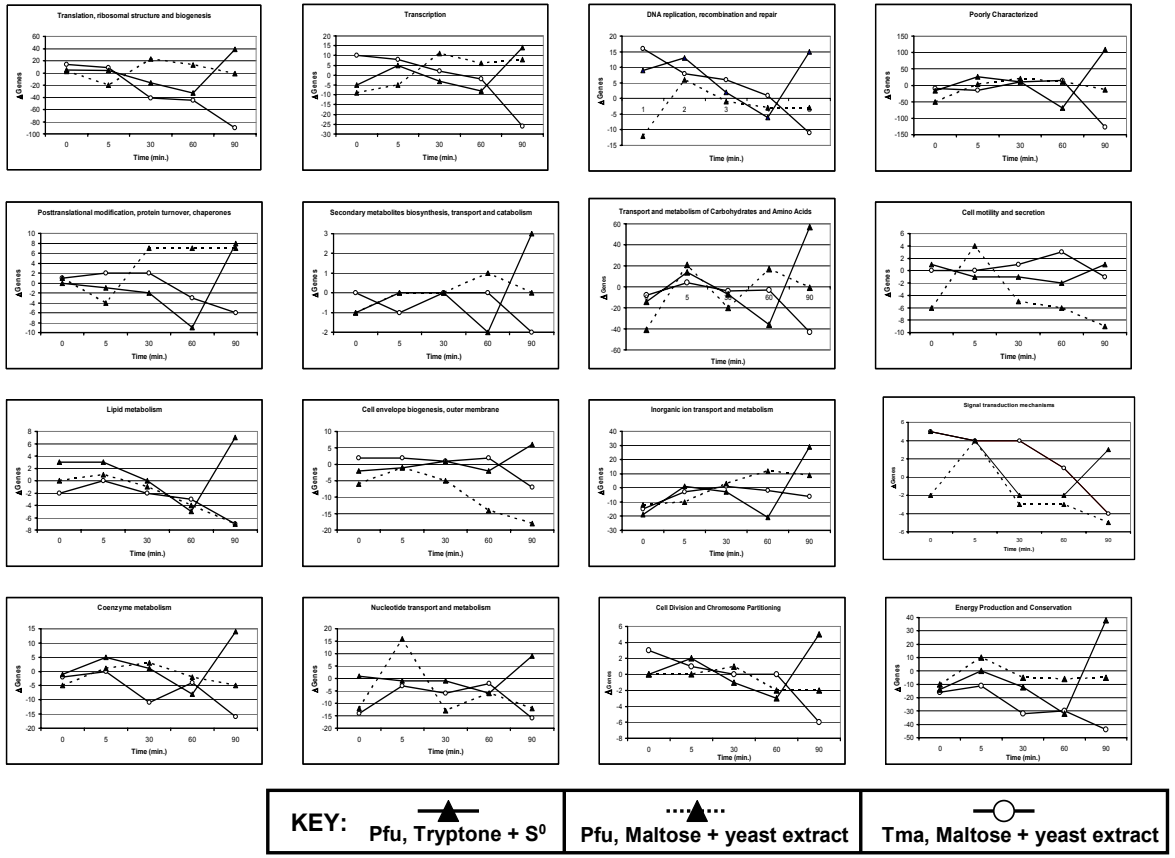


Figure 13. Changes in differentially expressed genes within functional categories. Shown are differences in the numbers of differentially expressed genes within a functional category (30, 31) over time in *P. furious* and *T. maritima*.

Chapter 6: Carbohydrate-Induced Differential Gene Expression Patterns in the Hyperthermophilic Archaeon *Pyrococcus furiosus*

*Keith R. Shockley, Shannon B. Conners, Clemente I. Montero, Matthew R. Johnson, Chung-Jung Chou, Stephanie L. Bridger, Nathan Wigner and Robert M. Kelly**

Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

To be Submitted to: ***Journal of Biological Chemistry (July, 2004)***

Running Title: *Pyrococcus furiosus* sugar utilization

*Address inquiries to: **Robert M. Kelly**
Department of Chemical Engineering
North Carolina State University
Raleigh, NC 27695-7905

Phone: (919) 515-6396
Fax: (919) 515-3465
Email: rmkelly@eos.ncsu.edu

Abstract

Full genome cDNA microarrays were used to assess differential gene expression patterns in the hyperthermophilic archaeon *Pyrococcus furiosus* grown on a range of glucose-based carbohydrates (i.e., maltose, cellobiose, barley glucan, laminarin, starch and chitin) and tryptone + S^0 as well as the effects of S^0 and yeast extract in the medium. About 80% of the annotated ORFs in the genome were differentially regulated ($P > 0.001$) between at least two different growth conditions. Gene expression analyses indicated that *P. furiosus* was able to recognize different sugar linkages and selectively regulate the transcription of processes involved in the regulation of sugar and peptide utilization processes. In addition, the effects of yeast extract and the presence of S^0 in the medium were found to influence gene expression patterns. The absence of yeast extract in the medium led to the induction of genes responsible for the biosynthesis of thiamin pyrophosphate, an essential cofactor in prokaryotic microorganisms. The transcriptional response data resulting from this study can be used to construct possible sugar utilization pathways in this organism, including the transport and processing of colloidal chitin. Taken together, the results presented here indicate that *P. furiosus* was able to adapt its transcriptional machinery to cope with changing environments.

Introduction

Pyrococcus furiosus is an obligately anaerobic, heterotrophic hyperthermophilic archaeon originally isolated from geothermal features associated with Vulcano Island, Italy (7). Most species of *Pyrococcus* can only use peptide-related substrates carbon sources and do not grow to significant levels without S^0. However, *P. furiosus* is able to utilize a range of poly- and oligosaccharide substrates for growth (2, 5, 11, 12). *P. furiosus* grows well at temperatures above 90^oC; once the organism reaches an optimal growth temperature it seeks metabolizable sources of carbon from its surroundings. Because the organism must constantly adapt its metabolism to the changing environment found in the open sea, it must be able to immediately recognize and accumulate important growth substrates from its environment, which may be present in very low concentrations. In order to utilize complex sources of carbon in their surroundings, microorganisms must be able to break down large substrate components into smaller components, transport those fragments into the cell, and hydrolyze them into glucose before using them for growth and maintenance. Yet, the transport of all proteins and sugars present in the environment costs much energy and uses valuable cell resources and many organisms are able to selectively adjust the expression of key metabolic genes in their genomes to accommodate such processes. In *P. furiosus*, glucose is converted into pyruvate through either the modified Embden-Meyerhof (EM) or Entner Doudoroff (ED) pathways (26).

The presence of multiple gene clusters in the *P. furiosus* genome sequence suggests that the organism is able to selectively utilize a range of glucose-based carbon sources. However, very little is known about the coordinated strategies for sugar transport in these organisms (12). The full genome sequence contains genes encoding glycoside hydrolases,

extracellular proteases and transporters. Microorganisms are known to transport sugars through three main transport systems: secondary transport, phosphoenolpyruvate-dependent phosphortransferase systems (PTS) and ATP binding cassette (ABC) transport (10). Biochemical studies suggest that the PTS systems are absent in the hyperthermophilic archaea and no secondary transporters have been implicated in sugar transport in the archaea (10). However, ABC transporters are abundant in the genome sequence and have been biochemically characterized for the uptake of cellobiose (11), maltodextrins (12) and trehalose/maltose (12). The ABC transporters are found in all domains of life, ranging from the gram negative bacteria to humans. ABC transporters are subdivided into the carbohydrate uptake transporter and the di/oligopeptide (Opp) transporter families (19), both of which are present in the genome sequence annotation of *P. furiosus*. It should be noted that a glucose transporter from *Solfolobus solfataricus* (6) and several sugar transporters from *Thermotoga maritima* ((3) and Conners et al., personal communication) have been misannotated in respective genome sequences as Opp transporters. From previous growth experiments and characterization of specific enzymes, *P. furiosus* is known to utilize both polysaccharides and simple sugars for growth, including cellobiose (11, 27), laminarin (27), chitin (8), maltose (12) and starch (12) as well as tryptone in the presence of elemental sulfur (22).

Regulation of genes encoding specific carbohydrate-active proteins in *P. furiosus* has only been studied to a limited extent thus far (20). Here, a full genome cDNA microarray was used in conjunction with mixed model analysis to explore saccharide and protein utilization by this organism. In addition to proteolytic growth, glucose-based linkages of various di- and poly-saccharides were examined, including β-1,3 (laminarin), β-1,4 (cellobiose), β-1,3 and β-1,4 (barley) and α-1,6 (maltose, starch) linkages.

Materials and methods

Array design

DNA primers were designed for all 2065 open reading frames in the genome sequence of *Pyrococcus furiosus* DSM 3638 based on sequence information found online at the site for The Institute for Genomic Research at (http://www.tigr.org/tigr-scripts/CMR2/GenomePage3.spl?database=ntpf01). DNA primers were designed with similar annealing temperatures and minimal hairpin formation using Genomax 2.0 (Informax, Bethesda, MD). All primers were ordered from IDTDNA. Probes were amplified in the PTC-100 Thermocycler (MJ Research, Inc., Waltham, MA) using Taq polymerase (Boehringer, Indianapolis, IN) and genomic DNA isolated from cell cultures according to previously published protocols (17). After amplification, the PCR products were quantitated and purified to a concentration of 100 ng/µl using QIAquick PCR purification kits (Qiagen, Valencia, CA). Purified PCR products were re-suspended in 50% DMSO, randomized, dispensed evenly into microarray printing plates (Genetix) and printed onto ULTRAGAPS aminosilane-coated microscope slides (Corning, Corning, NY) with a QArray-Mini Arrayer (Genetix, London, UK). The DNA was attached to the substrate by UV crosslinking in a GS GeneLinker UV Chamber (BioRad, Hercules, CA) at 250 mJ and then baked at 75^{o}C for 2 hours. The crosslinked slides were protected stored desiccated and away from light at room temperature until they were used.

Growth of Pyrococcus furiosus

Pyrococcus furiosus was cultured anaerobically at 90^{o}C on Sea Salts Medium (SSM) essentially as described previously (22), with a growth substrates (sugar, chitin or tryptone)

and 1 g/L yeast extract as primary carbon/energy sources. Growth on tryptone always included S^0, growth on cellobiose and maltose was performed with and without the presence of S^0, and no S^0 was present for growth on the remaining carbohydrates. Growth substrates maltose, laminarin, cellobiose, starch and chitin were obtained from Sigma (Sigma, St. Louis, MO). Tryptone was obtained from Fisher. All sugar-based substrates were added to SSM (final concentration 3.3 g/l) as a carbon source prior to inoculation. Growth on chitin was performed as described previously (8). A 60 ml batch culture was used to inoculate SSM medium supplemented with 3.3 g/l maltose in a 1 L pyrex bottle. *P. furiosus* was grown at 90^{o}C until early log phase, after which one ml of sample was removed for cell density enumeration by epifluorescent microscopy with acridine orange staining (9).

RNA isolation

Cultures were started with a 0.5% inoculum from a 60 mL pre-culture grown the previous night. To harvest, the cells were passed through a coffee filter to remove residual elemental sulfur and centrifuged at 7500 rpm (8145g) for 22 min. at 4^{o}C. Cells were then re-suspended in ice-cold SSM, aliquotted into 2.0 mL eppendorf tubes (Ambion) and centrifuged at 13,800g for 30 sec. The resulting pellet was re-suspended in 85 μL of cold TE buffer, and then ruptured in 625 μL of lysis buffer (50 mM glucose, 10 mM EDTA, 25 mM Tris). The lysate was passed through a 20-gauge needle to shear the genomic DNA, after which 62.5 mL of 2 M sodium acetate (pH 5.2) was added to the resulting mixture. Then, an equal volume of acidic phenol/chloroform (5:1) was added, the aqueous phase was extracted, and the RNA was ethanol-precipitated overnight at –20^{o}C. After washing with 70% ethanol, the RNA pellet was re-suspended in 10 mM Tris (pH 8.5) and passed through fiber filter

columns from the RNAqeuous™ RNA isolation kit (Ambion). Integrity of the RNA was confirmed by visual inspection on 1.0% native agarose gels as well as by measuring the A_{260}/A_{280} ratio with a DU® 640 Spectrophotomer (Beckman Coulter, Inc., Fullerton, CA).

Generation of cDNA and hybridization

First-strand cDNA was prepared from total RNA using Stratascript (Stratagene) and random hexamer primers (Invitrogen Life Technologies, Carlsbad, CA) by the incorporation of 5-[3-Aminoallyl]-2'-deoxyuridine-5'-triphosphate (aa-dUTP) (Sigma) as per (22). Hybridizations and washes were performed as described previously (22). The slides were scanned using the Scanarray 4000 scanner (Perkin Elmer, Fremont, CA). Signal intensity data for all experiments was extracted using Scanarray (Perkin Elmer).

Mixed model analyses

Replication of treatments, arrays, dyes, and cDNA spots allowed the use of analysis of variance (ANOVA) models for data analysis (28). Loop designs were constructed as indicated in the text and reciprocal labeling was utilized for all samples so that dye effects could be estimated. Spot intensities obtained from Scanarray (Perkin Elmer) and imported directly into SAS (SAS Institute, Cary, NC). A linear normalization ANOVA model (28) was used to estimate global variation in the form of fixed (dye (D), treatment (T)) and random (array (A), block (B) and spot (S)) effects and random error using the model $\log_2(y_{ijklmn}) = A_i + D_j + T_k + A_i(B_lS_m) + A_i(B_l) + \varepsilon_{ijklm}$. A gene-specific ANOVA model was then used to partition the remaining variation into gene-specific effects using the model $r_{ijklmn} = \mu + A_i + D_j + T_k + A_i(B_lS_m) + A_i(B_l) + \gamma_{ijklm}$.

For complete information on significance of expression changes, fold changes, pairwise volcano plots, and hierarchical clustering for all of the genes included on the array, see our website (follow the microarray link, data to be posted upon acceptance of the manuscript) at: http:///www.che.ncsu.edu/extremophiles/.

Results and discussion

Experimental layout

A full genome cDNA microarray was constructed to include all 2065 annotated ORFs present in the genome sequence of the heterotrophic hyperthermophile *Pyrococcus furiosus* (15). A total of 2015 ORFs yielded at least one detectable fluorescence intensity for at least one treatment condition. Of particular interest were genes encoding proteins related to the transport and metabolism of saccharides and peptides. These genes are evident from COG functional categories (25) and extensive work on Glycoside Hydrolases by Coutinho and Henrisat (4) available at online at (http://afmb.cnrs-mrs.fr/~cazy/CAZY.index.html). *P. furiosus* was grown in the presence of various glucose-based carbon/energy sources at a growth temperature of 90^{o}C (see Figure 1). At least 6 passes were performed with a 0.5% (w/v) inoculation before growth curves were generated; negative controls containing 1 g/L of yeast extract without the growth substrate were inoculated in parallel in order to ensure that observed growth was due to the substrate of interest. In most cases, cell densities reached greater than 10^8 cells/mL for studied growth conditions while no significant growth was observed for control cultures. As shown in Figure 2, a loop design was constructed in order to analyze the transcriptional profiles of *P. furiosus* grown with various carbon/energy sources provided in the medium.

In addition to growth using tryptic digests, the effect of growth on sugars with various glucose-based linkages of different di- and poly-saccharides were examined, including β-1,3 (laminarin), β-1,4 (cellobiose, chitin), β-1,3 and β-1,4 (barley), β-1,6 (maltose, starch) linkages. As illustrated in Figure 2, *P. furiosus* had the fastest doubling times for growth on tryptone and elemental sulfur when yeast extract was available in the medium (t_D = 46 min.),

but also grew rapidly on barley (t_D = 49 min.), maltose (t_D = 54 min.) and starch (t_D = 56 min.). The slowest growth occurred when metabolizing cellobiose (t_D = 80 min.).

Trends in carbohydrate utilization

Figure 3 shows overall least square mean and standardized least square mean heat plots that summarize the expression patterns of all 2015 ORFs measured on the *P. furiosus* array during growth on all growth conditions. The least square means cluster compares the normalized expression levels of each gene in the organism within each treatment. Table 1 lists genes encoding known and putative glycosyl hydrolases based on family classification algorithms derived from amino acid sequence similarities [tool available online at (4) (http://afmb.cnrs-mrs.fr/~cazy/CAZY/index.html)], along with an indication of which growth substrates induced the expression of the ORF based on the level of expression given by least square mean estimates of treatment effects at a level of $\log_2 R > 0.2$. A total of 5 ORFs are predicted to encode extracellular glycosyl hydrolases; PF0854, PF0076, PF0477, PF1234, and PF1935. Results here were consistent with previous results in which the gene encoding a laminarinase (LamA) was induced when *P. furiosus* was grown on cellobiose and laminarin, but not on maltose (27) and a previous study implicating PF1234, an extracellular chitinase, in the utilization of chitin in *P. furiosus* (8). In addition, the gene encoding an exported alpha amylase (PF1935) was induced for growth on maltose, but not tryptone, as found previously (20). Significant expression of each gene predicted to encode an exported enzyme, based on the presence of a putative signal peptide, is induced by at least one growth condition.

The standardized least square means cluster shown in Figure 1 compares the expression levels of each ORF across treatment conditions. Here, as with a previous study

investigating carbohydrate utilization in *Thermotoga maritima* (3), it was advantageous to view standardized clusters in order to discover similar expression patterns of genes relative to other treatments in order to discern treatment-specific regulation patterns (shown later in Fig. 5, 6 and 7).

Effect of yeast extract

Figure 4 shows a volcano plot in which least square means (LSMs) differences between *P. furiosus* grown on tryptone + S^0 without yeast extract vs. growth on tryptone + S^0 with yeast extract are plotted against a $-\log_{10}$(P-value) of the difference. As illustrated in Figure 4, many transcriptional changes appear to result from the addition of yeast extract to the medium which indicates that complex media may have a significant effect on differential gene expression. In Figure 4, LSM differences greater than 2 (analogous to fold changes greater than 4-fold) have highly significant comparisons (P-value $\leq 10^{-15}$). As shown in the figure, 10 ORFs have significantly higher expression for growth on tryptone without yeast extract while only 1 ORF, a hypothetical protein (PF0730) has a higher expression during growth with yeast extract in the medium under this threshold criterion (4-fold change, $P < 10^{-15}$). Thiamin pyrophosphate biosynthesis appears to be more important without yeast extract in the medium, as genes encoding a thiamin biosynthetic enzyme (PF1530), a phosphomethylpyrimidine kinase [hmp-phosphate kinase] (PF1333), a hydroxyethylthiazole kinase (PF1335), a thiamin phosphate phosphorylase (PF1338) and a phosphoribosylaminoimidazole carboxylase (PF0426) were expressed at elevated levels on growth without yeast extract in the medium. Each of the enzymes encoded by these genes have been implicated in thiamin pyrophosphate biosynthesis, an essential coenzyme in

prokaryotes (16). Two transcriptional activators (PF1337 and PF1338) were expressed at significantly higher levels without yeast extract in the medium which, judging by their similar gene expression pattern and proximity to the thiamin biosynthesis ORFs, may regulate the biosynthesis of thiamin in *P. furiosus*.

Influence of elemental sulfur in the growth medium

P. furiosus is one of the few hyperthermophiles that can grown in the absence of sulfur and has therefore been the focus of many of the studies conducted on sulfur metabolism in high temperature organisms (18, 21, 23). *P. furiosus* will not grow to significant levels on peptide-based substrates without the addition of S^0 (1); however, the mechanism of the reduction of S^0 to H_2S is not known. A previous study, investigating the expression of 271 ORFs in *P. furiosus* grown on maltose with or without sulfur found that expression levels associated with sugar and peptide catabolism, the metabolism of metals and the biosynthesis of various cofactors, amino acids and nucleotides showed increased expression with the addition of S^0 while three different hydrogenase systems were down-regulated with S^0 in the medium (21). Here, for growth on maltose, all significant changes in transcription (P-value < 0.01) confirmed the results from (21). However, the previous analysis was conducted with a limited number of ORFs and differential gene expression was based on large fold change differences unnormalized for effects of dye and array, which can show significant effects (see Chapter 3).

Here, a mixed model analysis of full genome microarray data and was conducted with two different growth substrates (cellobiose and maltose) for comparison. The presence of elemental sulfur in the growth medium influences the transcriptional profile of *P. furiosus*

when the organism is grown on cellobiose or maltose (Figure 5). A total of 8 ORFs were expressed at higher levels without elemental sulfur in the medium when *P. furiosus* was grown on maltose, while 13 ORFs were expressed higher on cellobiose growth without sulfur in the medium for threshold levels of $\log_2$(differences) > 2.0 and P-value < 6.0 x 10^{-7}. Interestingly, only 2 of the ORFs most dramatically regulated in the previous study on maltose were amongst the most significantly down-regulated ORFs here; PF0893 encodes a subunit of a previously characterized hydrogenase (13) and PF1428 may encode a subunit of a membrane bound hydrogenase identified by Schut et al. (21) in a previous study. Only one ORF from the most dramatically up-regulated genes previously (21) were among the most up-regulated genes here, PF2025. For the same threshold criteria, 4 ORFs were expressed higher with elemental sulfur in the medium when *P. furiosus* was grown on maltose and 10 ORFs were expressed at higher levels for growth on cellobiose with S^0 present for both growth substrates, a ribonucleotide reductase PF0440. The only gene that is expressed higher on both sugars without sulfur in the medium with the indicated cutoff levels was an ORF encoding a hypothetical protein, PF0924 (hypothetical protein).

Growth on substrates containing α-1,6 sugar backbone linkages

Standardized LSM clusters C2 and C3 in Figure 3 were expanded in Figure 6. These clusters show sugar utilization patterns specific for growth on maltose and starch, which contain α-1,6- sugar backbone linkages. Genes from annotated maltose/trehalose (PF1739-1744) and maltodextrin operons (PF1933-1936, PF1938-PF1939) were induced. Previous studies have shown that genes from both operons have been shown to be differentially expressed when *P. furiosus* is grown on maltose (12, 20) or starch (12), even though the

maltodextrin operon only transports maltotriose or higher maltooligosaccharides but not maltose (12). Here, additional components of pathways responsible for the utilization of α-1,6-linked sugars were identified, including genes encoding additional components of transporters (PF0905, PF0744), processing enzymes (PF0272, PF1535), enzymes involved in downstream processing of cleavage products (PF1784, PF1344) and numerous hypothetical proteins warranting further biochemical investigation.

Growth on substrates containing β-1,4- and β-1,3- sugar backbone linkages

Figure 7 shows STLSM plots of genes encoding enzymes in the utilization of sugars containing β-1,4- and β-1,3- sugar backbone linkages. As determined previously, genes encoding CelB, LamA and two alcohol dehydrogenases were expressed at high levels for growth on cellobiose and laminarin (27). In addition, the gene encoding *ctbA*, a cellobiose binding protein previously studied (11), was expressed at moderate levels for growth on cellobiose (both with and without elemental sulfur in the medium) and laminarin (data not shown). Here, a possible additional transport system involving PF1696 and PF1697 was identified.

Chitin utilization

A strong induction of ORFs involved in chitin utilization was evident from standardized least square means clusters (Figure 8). In particular, gene clusters involving homologs to chitinase processing genes (PF0354-PF0356) as per a previous study conducted with *Thermococcus kodakaraensis* KOD1 (24) and putative ABC transporters PF0357-PF0361, showed strong gene expression when *P. furiosus* was grown on chitin. In addition,

genes encoding an extracellular chitinase (PF1234) and an intracellular chitinase (PF1233) shown to be able to hydrolyze chitin (8) were induced. The strong induction of these ORFs enabled a possible construction of a pathway for chitin utilization in *P. furiosus* (Figure 9). In particular, the extracellular chitinase (PF1234) may degrade extracellular chitin, which can be transported into the cell via the PF0357-PF0361 ABC transport system annotated as an oligopeptide (Opp) transport system. Once inside the cell, $GlcNAc_n$ fragments can be hydrolyzed to $GlcNAc_2$ through PF1233 and eventually processed to GlcN-GlcN through the action of a homolog to the *T. kodakaraensis* KOD1 deacetylase (24), PF0354. A homolog to the *T. kodakaraensis* KOD1 glucosaminidase (24), PF0363, may process GlcN-GlcN to GlcN, which can then be phosphorylated (PF0356) and isomerized (PF0362) to a metabolizable sugar entering glycolysis.

Regulation of ABC transporters

In addition to the gene expression here that challenges genome sequence annotation of Opp family regulators in hyperthermophiles, studies in other hyperthermophilic organisms have found that ABC transporters have not been annotated correctly (3, 6). To further investigate the annotation of ABC transporters in the genome sequence, all annotated sugar and di/oligopeptide transporters were separated from the full data set and clustered based on LSM and STLSM estimates (Figure 10, I and II). Results indicated that the expression of ABC transporters was specific to substrate (Figure 10, II). For instance, PF0116-PF0119 was strongly induced for growth on barley; PF1696-PF1697 were most strongly induced on cellobiose and laminarin; PF1933-PF1939 were most strongly up-regulated for growth on maltose and starch; PF1739-PF1747 showed strong induction for growth on maltose, starch,

laminarin and chitin; PF0357-PF0361 were most strongly expressed for growth on chitin and PF0191-PF0194 were strongly induced for growth on tryptone and sulfur.

Concluding remarks

The obligately anaerobic heterotrophic hyperthermophilic archaeon *P. furiosus* was grown on a variety of carbon/energy sources and shown to be able to selectively regulate the expression of its genetic inventory in response to the presence of these substrates in the medium. Previous transcriptional work studying cellobiose, laminarin, maltose, and starch was confirmed and the full genome analysis led to the identification of many additional components to the substrate utilization pathways in this organism. Taken together, the results suggest that this high temperature organism is able to selectively recognize differences in sugar backbone linkages of glucose-based saccharides and recruit carbohydrates from the medium to respond to environmental changes. In addition, *P. furiosus* is able to modify its transcriptional profile to adjust to the presence or absence of yeast extract in the medium as well as the addition of elemental sulfur.

References

1. **Adams, M. W. W., J. F. Holden, A. L. Menon, G. Schut, A. M. Grunden, C. Hou, A. M. Hutchins, J. Jenny, F. E., C. Kim, K. Ma, G. Pan, R. Roy, R. Sapra, S. V. Story, and M. F. Verhagen.** 2001. Key role for sulfur in peptide metabolism and in regulation of three hydrogenases in the hyperthermophilic archaeon *Pyrococcus furiosus*. J. Bacteriol. **183:**716-724.

2. **Bauer, M. W., S. B. Halio, and R. M. Kelly.** 1996. Proteases and glycosyl hydrolases from hyperthermophilic microorganisms. Adv. Prot. Chem. **48:**271-310.

3. **Chhabra, S. R., K. R. Shockley, S. B. Conners, K. L. Scott, R. D. Wolfinger, and R. M. Kelly.** 2003. Carbohydrate-induced differential gene expression patterns in the hyperthermophilic bacterium *Thermotoga maritima*. J. Biol. Chem. **278:**7540-7552.

4. **Coutinho, P. M., and B. Henrissat.** 1999. Carbohydrate-active enzymes: an integrated database approach, p. 3-12. *In* H. J. Gilbert, G. Davies, B. Henrissat, and B. Svensson (ed.), Recent Advances in Carbohydrate Bioengineering. The Royal Society of Chemistry, Cambridge.

5. **Driskill, L. E., K. Kusy, M. W. Bauer, and R. M. Kelly.** 1999. Relationship between glycosyl hydrolase inventory and growth physiology of the hyperthermophile *Pyrococcus furiosus* on carbohydrate-based media. J. Bacteriol. **65:**893-897.

6. **Elferink, M. G., S. V. Albers, W. N. Konings, and A. J. Driessen.** 2001. Sugar transport in *Sulfolobus solfataricus* is mediated by two families of binding protein-dependent ABC transporters. Mol Microbiol **39:**1494-1503.

7. **Fiala, G., and K. O. Stetter.** 1986. *Pyrococcus furiosus* sp. nov. represents a novel genus of marine heterotrophic archaebacteria growing optimally at 100^{o}C. Arch. Microbiol. **145:**56-61.

8. **Gao, J., M. W. Bauer, K. R. Shockley, M. A. Pysz, and R. M. Kelly.** 2003. Hyperthermophilic archaeon *Pyrococcus furiosus* growth on chitin involves two family 18 chitinases. Appl. Environ. Microbiol. **69:**3119-3128.

9. **Hobbie, J. E., R. J. Daley, and S. Jasper.** 1977. Use of nuclepore filters for counting bacteria by fluorescence microscopy. Appl. Environ. Microbiol. **33:**1225-1228.

10. **Koning, S. M., S. V. Albers, W. N. Konings, and A. J. Driessen.** 2002. Sugar transport in the (hyper-) thermophilic archaea. Res. Microbiol. **153**:61-67.

11. **Koning, S. M., M. G. Elferink, W. N. Konings, and A. J. Driessen.** 2001. Cellobiose uptake in the hyperthermophilic archaeon *Pyrococcus furiosus* is mediated by an inducible, high-affinity ABC transporter. J. Bacteriol. **183:**4979-4984.

12. **Koning, S. M., W. N. Konings, and A. J. M. Driessen.** 2002. Biochemical evidence for the presence of two a-glucosidase ABC-transport systems in the hyperthermophilic archaeon *Pyrococcus furiosus*. Archaea **1:**19-25.

13. **Ma, K., R. N. Schicho, R. M. Kelly, and M. W. W. Adams.** 1993. Hydrogenase of the hyperthermophile *Pyrococcus furiosus* is an elemental sulfur reductase or sulfhydrogenase: evidence for a sulfur-reducing hydrogenase ancestor. Proc. Natl. Acad. Sci. **90:**5341-5344.

14. **Nielsen, J., J. Englebrecht, S. Brunak, and G. von Heijne.** 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Prot. Eng. **10:**1-6.

15. **Robb, F. T., D. L. Maeder, J. R. Brown, J. DiRuggiero, M. D. Stump, R. K. Yeh, R. B. Weiss, and D. M. Dunn.** 2001. Genomic sequence of hyperthermophile, *Pyrococcus furiosus*: implications for physiology and enzymology. Meth. Enzymol. **330:**134-157.

16. **Rodionov, D. A., A. G. Vitreschak, A. A. Mironov, and M. S. Gelfand.** 2002. Comparative genomics of thiamin biosynthesis in procaryotes. J. Biol. Chem. **277:**48949-48959.

17. **Sambrook, J., E. H. Fritsh, and T. Maniatis.** 1989. Extraction, purification, and analysis of messenger RNA from eukarytoic cells. Cold Spring Harbor Laboratory Press, Plainview, NY.

18. **Schicho, R. N., K. Ma, M. W. W. Adams, and R. M. Kelly.** 1993. Bioenergetics of sulfur reduction in the hyperthermophilic archaeon *Pyrococcus furiosus*. J. Bacteriol. **175:**1823-1830.

19. **Schneider, E.** 2001. ABC transporters catalyzing carbohydrate uptake. Res. Microbiol. **152:**303-310.

20. **Schut, G. J., S. D. Brehm, S. Datta, and M. W. Adams.** 2003. Whole-genome DNA microarray analysis of a hyperthermophile and an archaeon: *Pyrococcus furiosus* grown on carbohydrates or peptides. J. Bacteriol. **185:**3935-3947.

21. **Schut, G. J., J. Zhou, and M. W. Adams.** 2001. DNA microarray analysis of the hyperthermophilic archaeon *Pyrococcus furiosus*: evidence for anNew type of sulfur-reducing enzyme complex. J. Bacteriol. **183:**7027-7036.

22. **Shockley, K. R., D. E. Ward, S. R. Chhabra, S. B. Conners, C. I. Montero, and R. M. Kelly.** 2003. Heat shock response by the hyperthermophilic archaeon *Pyrococcus furiosus*. Appl. Environ. Microbiol. **69:**2365-2371.

23. **Stetter, K. O.** 1996. Hyperthermophilic procaryotes. FEMS Microbiol. Rev. **18:**149-158.

24. **Tanaka, T., T. Fukui, H. Atomii, and T. Imanaka.** 2003. Characterization of an exo-*b*-D-glucosamine involved in a novel chitinolytic pathway from the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1. J Bacteriol **185:**5175-5181.

25. **Tatusov, R. L., E. V. Koonin, and D. J. Lipman.** 1997. A genomic perspective on protein families. Science **278:**631-637.

26. **Verhees, C. H., S. W. M. Kengen, J. E. Tuininga, G. J. Schut, M. W. W. Adams, W. M. de Vos, and J. van der Oost.** 2003. The unique features of glycolytic pathways in archaea. Biochem. J. **375:**231-246.

27. **Voorhorst, W. G. B., Y. Guegen, C. M. Geerling, G. Schut, I. Dahlke, M. Thomm, J. van der Oost, and W. M. de Vos.** 1999. Transcriptional regulation in the hyperthermophilic archaeon *Pyrococcus furiosus*: coordinated expression of divergently oriented genes in response to b-linked glucose polymers. J. Bacteriol. **181:**3777-3783.

28. **Wolfinger, R. D., G. Gibson, E. D. Wolfinger, L. Bennett, H. Hamadeh, P. Bushel, C. Afshari, and R. S. Paules.** 2001. Assessing gene significance from cDNA microarray expression data via mixed models. J. Comput. Biol. **8:**625-637.

Table 1. Predicted Signal Peptides in Glycosidases from *Pyrococcus furiosus*

Locus	Known or Putative Activity	Microarray *	Signal Peptide	Sequence[+]	Cleavage Site (Amino acid#)
ENDO- Acting					
PF0854	β-1,4-Glucanase (Cel12)	---	Y	MSKKKFVIVSILTILLVQA ^ IY	19-20, 23-24, 19-20
PF0076	β-1,3-Glucanase (Lam16)	BA, CB, CB/S, LA, TY	Y	MKKEALLFLSLIFLVFVS ^ GCIHHST	19-20, 22-23, 23-24
PF0477	α-Amylase (Amy13)	---	Y	MNIKKLTPLLTLLLFFIVLASPVSA ^ AK	25-26, 25-26, 25-26
PF1234	Chitinase (Chi18A)	CH, TY	Y	MKTRMLGIVLAWLVVLSLVSPTISLFYPVSA ^ QQTV	31-32, 31-32, 31-32
PF1935	Pullulanase (Amy57B)	CB, MA, MA/S, ST	Y	MSRKLSLLLVFLIFGSMLG ^ ANNIVKA	19-20, 26-27, 19-20
PF1233	Chitinase (Chi18B)	CH, T/S	N	--	--
PF0272	α-Amylase (Amy57A)	MA, MA/S, ST	N	--	--
PF1939	Neopullulanase (ORF13)	BA, CB, CH, MA, MA/S, ST	N	--	--
PF0478	ORF (ORF13)	---	N	--	--

Table 1 (cont.)

EXO- Acting					
PF0073	β-Glucosidase (Cel1)	CB, CB/S	N	--	--
PF0442	β-Glucosidase (ORF1)	CH, T	N	--	--
PF0356	ORF (ORF1)	BA, CH	N	--	--
PF1208	β-Mannosidase (Man1)	BA, CB, CB/S, LA, MA/S, T, T/S	N	--	--
PF0444	α-galactosidase (Gal57)	BA, CH, MA, T, TY	N	--	--
PF0363	ORF (ORF35)	BA, CB, CH, CB/S, LA, MA, MA/S, ST, T, TY	N	--	--

[+]**Algorithm :** SignalP V1.1 (http://genome.cbs.dtu.dk/services/SignalP/)
Networks trained on sequences from **gram-negative prokaryotes**, **gram-positive prokaryotes** and **eukaryotes** were used (14).
*refers to $\log_s R$ values of least square means estimates > 0.2. Key: BA, barley; CB, cellobiose; CH, chitin; CB/S, cellobiose + S^0, LA, laminarin; MA, maltose; MA/S, maltose + S^0; T, trypone + S^0; TY, tryptone + S^0 + yeast extract.

Figures

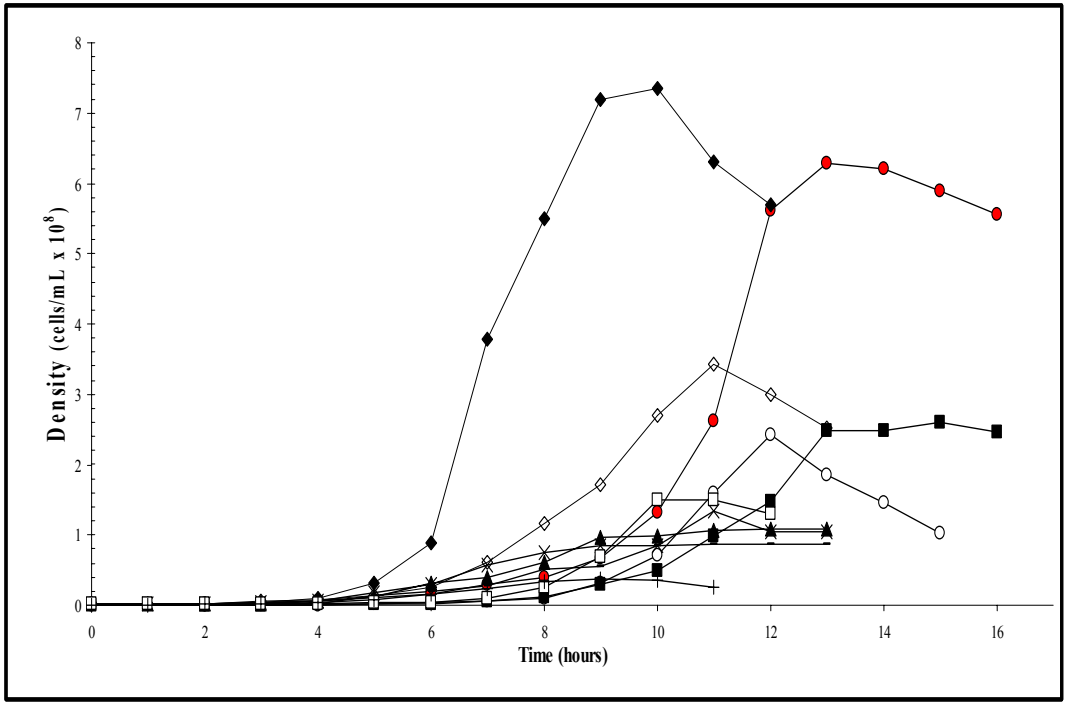


Figure 1. Growth curves for *P. furiosus* on various substrates added at initial concentrations of 3.3 g/L. Symbols: ●, cellobiose; ○, cellobiose + S^0; ■, maltose; □, maltose + S^0; -, barley β-glucan; **x**, starch; ▲, laminarin; +, chitin; ♦, tryptone + S^0 (with yeast extract); ◊, tryptone + S^0 (without yeast extract)

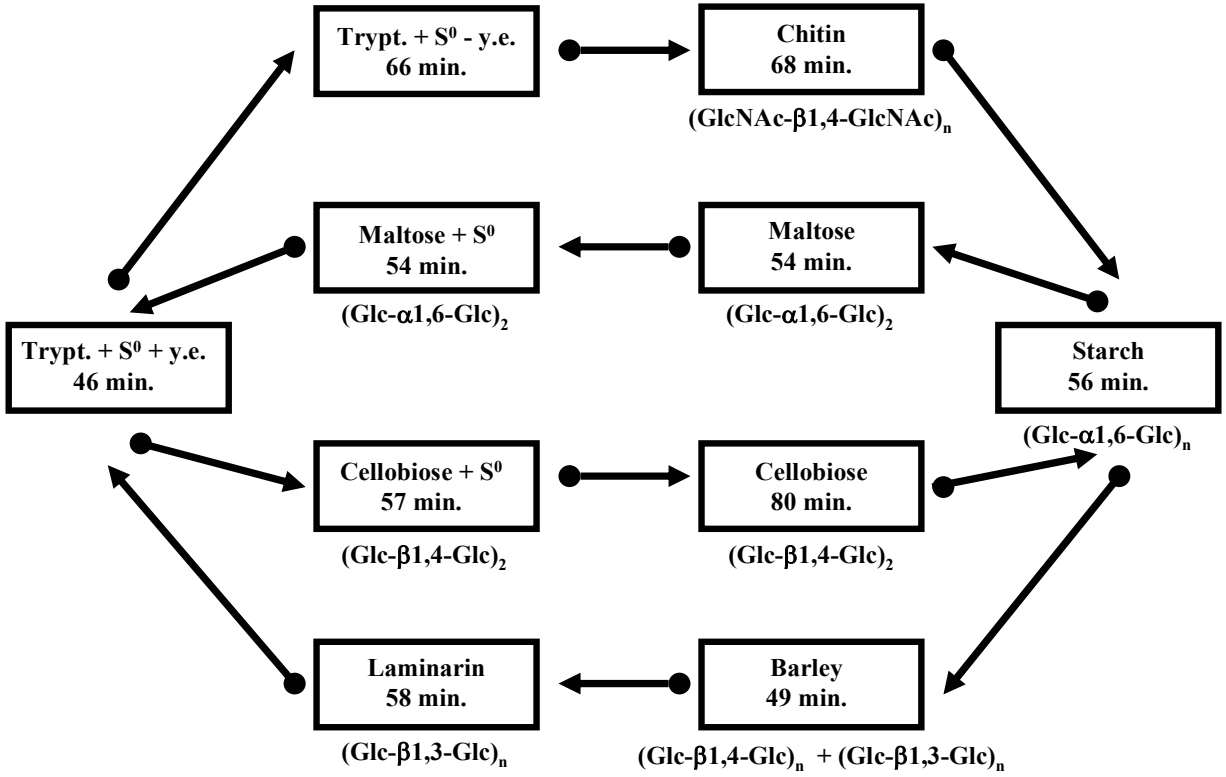


Figure 2. A-Optimal Loop Design (n=10 treatments; v=12 arrays) for the study of carbon-source utilization in *Pyrococcus furiosus.* The arrow ends correspond to the Cy3 and Cy5 channels as follows: Cy3 → Cy5. Doubling times are shown under the substrate name for each carbon/energy source listed.

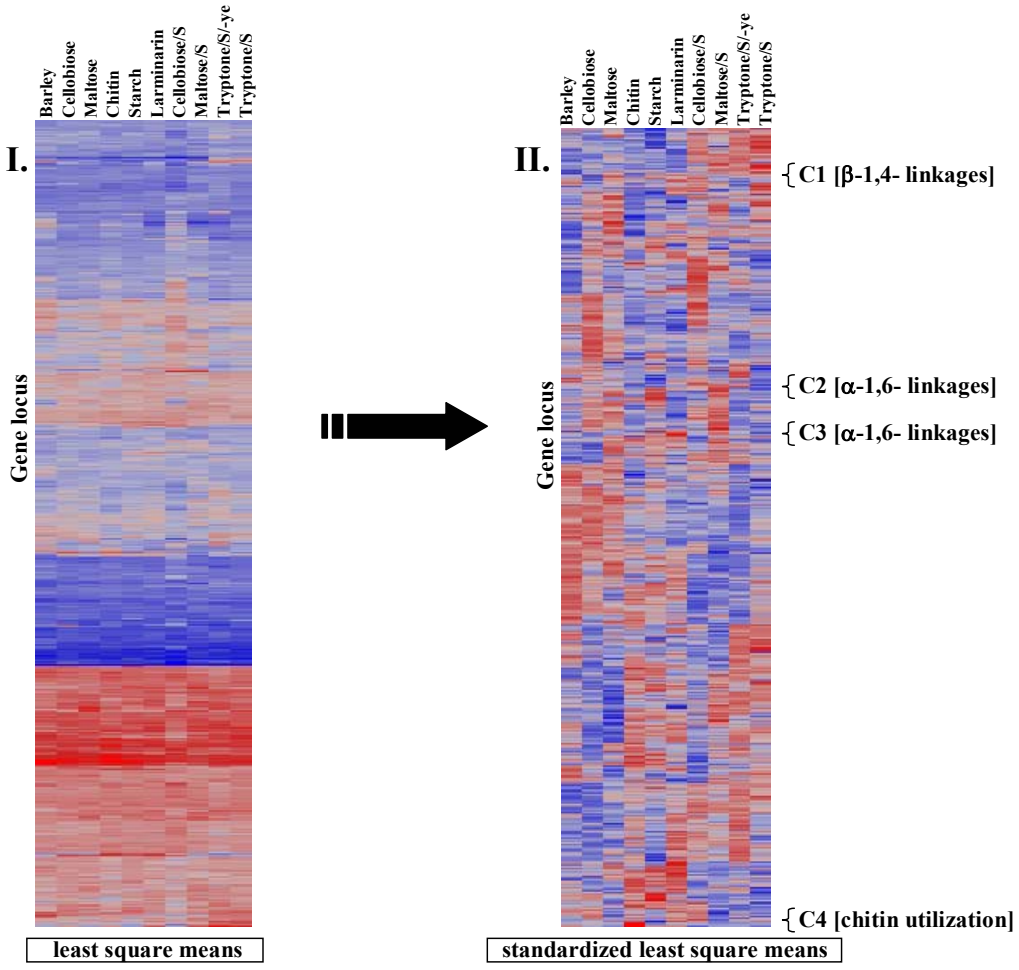
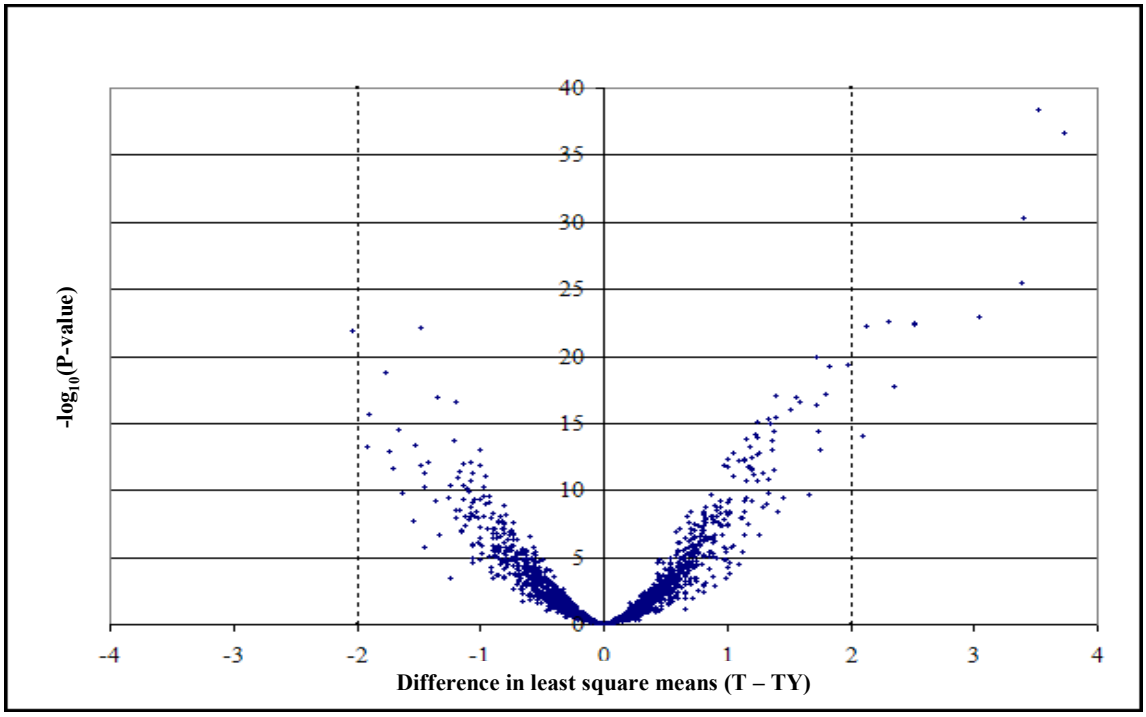
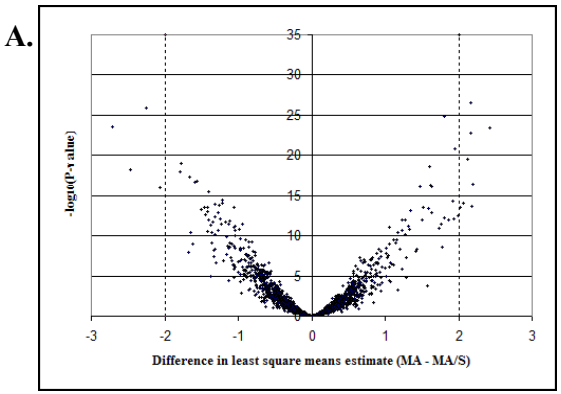


Figure 3. Hierarchical clusters constructed using least square means (I) and standardized least square means (II). Growth substates are displayed at the top of each cluster, and gene loci are presented along rows. Indicated clusters (C1 – C4) of the standardized least square means are shown in greater detail in Fig. 5 and 6.



+ PF0426, phosphoribosylaminoimidazole carboxylase
+ PF1333, phosphomethylpyrimidine kinase (hmp-phosphate kinase)
+ PF1334, thiamine phosphate pyrophosphorylase
+ PF1335, hydroxyethylthiazole kinase
+ PF1337, transcriptional activator, putative
+ PF1338, transcriptional activator, putative
+ PF1530, thiamine biosynthetic enzyme
+ PF1557, hypothetical protein
+ PF1602, glutamate dehydrogenase
+ PF1751, putative solute binding lipoprotein
- PF0730, hypothetical protein

Figure 4. Effect of including yeast extract in medium. The given volcano plot shows the $-\log_{10}$(P-value) plotted against differences in least square means (LSM) estimates of treatment effects for *P. furiosus* grown on tryptone + S^0 [T] and tryptone + S^0 + yeast extract [TY]. Dotted lines represent LSM differences of 2, which is analogous to 2^2 = 4-fold change in gene expression for a given ORF.



MA – MA/S

Locus	Description
up-regulated by maltose	
PF0126	DNA repair protein rad25
PF0729	multi domain protein containing corrinoid/iron-sulfur region
PF0730	hypothetical protein
PF0893	sulfhydrogenase delta subunit
PF0924	hypothetical protein
PF1428	hypothetical protein
PF1456	hypothetical protein
PF1573	TLDD protein homolog
up-regulated by S^0	
PF0440	ribonucleotide reductase
PF1761	hypothetical protein
PF1920	triosephosphate isomerase
PF1951	partial asparaginyl-tRNA synthetase matches COOH terminus

CB – CB/S

up-regulated by cellobiose	
PF0734	hypothetical protein
PF0909	ferric enterobactin transport ATP-binding protein homolog
PF0923	hypothetical protein
PF0924	hypothetical protein
PF1006	putative phosphate transport system permease protein C
PF1007	phosphate transport system permease psta
PF1527	hypothetical protein
PF1564	DNA-directed RNA polymerase subunit b
PF1704	hypothetical protein
PF1705	tryptophan synthetase alpha subunit
PF1739	trehalose/maltose binding protein
PF1742	putative trehalose synthase
PF1748	putative sulfate transport system permease protein, ABC transporter
up-regulated by S^0	
PF0043	phosphoenolpyruvate synthase (pyruvate, water dikinase)
PF0115	ATP-dependent 26S protease regulatory subunit
PF0440	ribonucleotide reductase
PF0456	carboxypeptidase I
PF1076	TLDD protein homolog
PF1077	hypothetical protein
PF1238	putative ABC transporter (ATP binding protein)
PF1421	hypothetical 4-aminobutyrate aminotransferase
PF1547	endoglucanase
PF2025	hypothetical protein

Figure 5. Effect of including elemental sulfur in the growth medium. The given volcano plot shows the $-\log_{10}$ (P-value) plotted against differences in least square means (LSM) estimates of treatment effects for *P. furiosus* grown on (A) cellobiose [CB] and cellobiose + S^0 [CB/S] and (B) maltose [MA] vs. maltose + S^0 [MA/S]. Dotted lines represent LSM differences of 2, which is analogous to 2^2 = 4-fold change in gene expression for a given ORF.

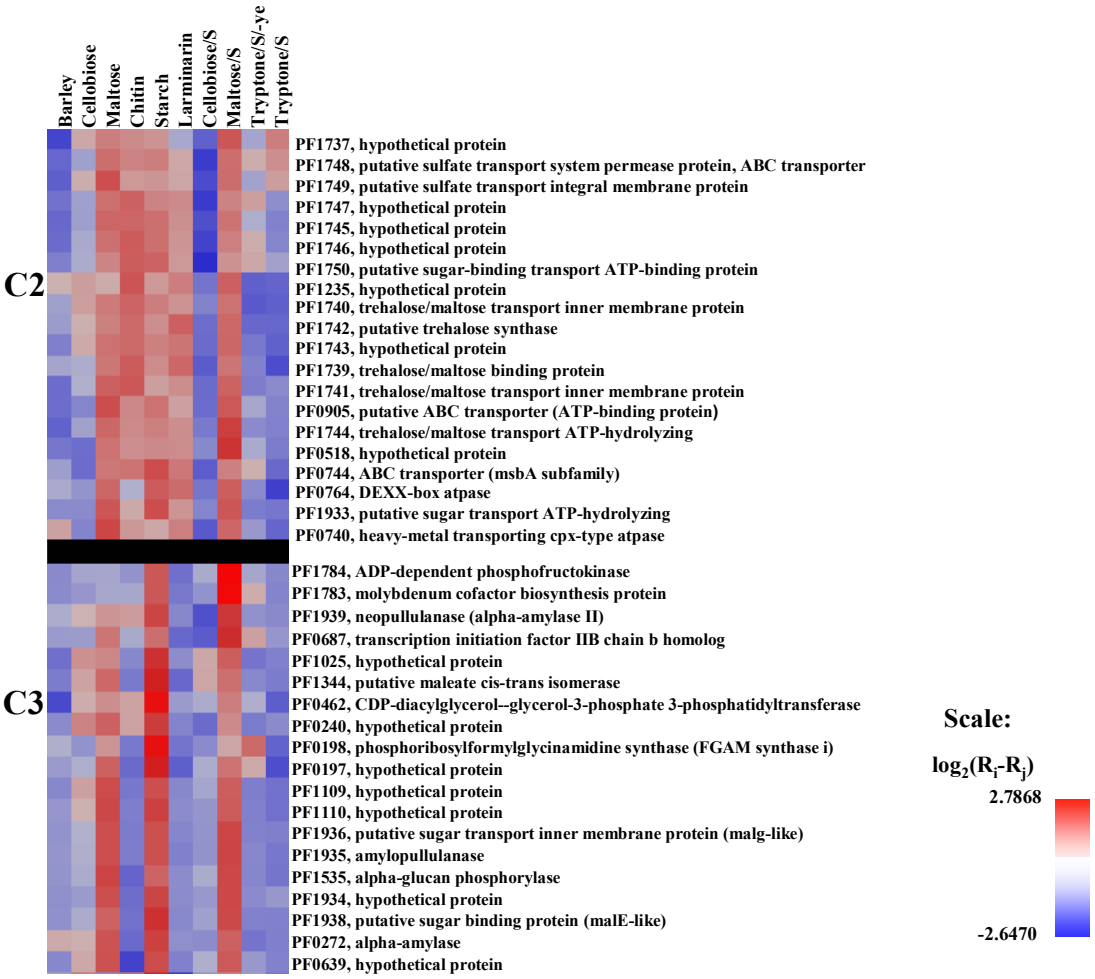


Figure 6. Maltose and starch dependent regulation. Sample clusters are constructed with standardized least square means estimates of treatment effects. Known or putative functions are indicated.

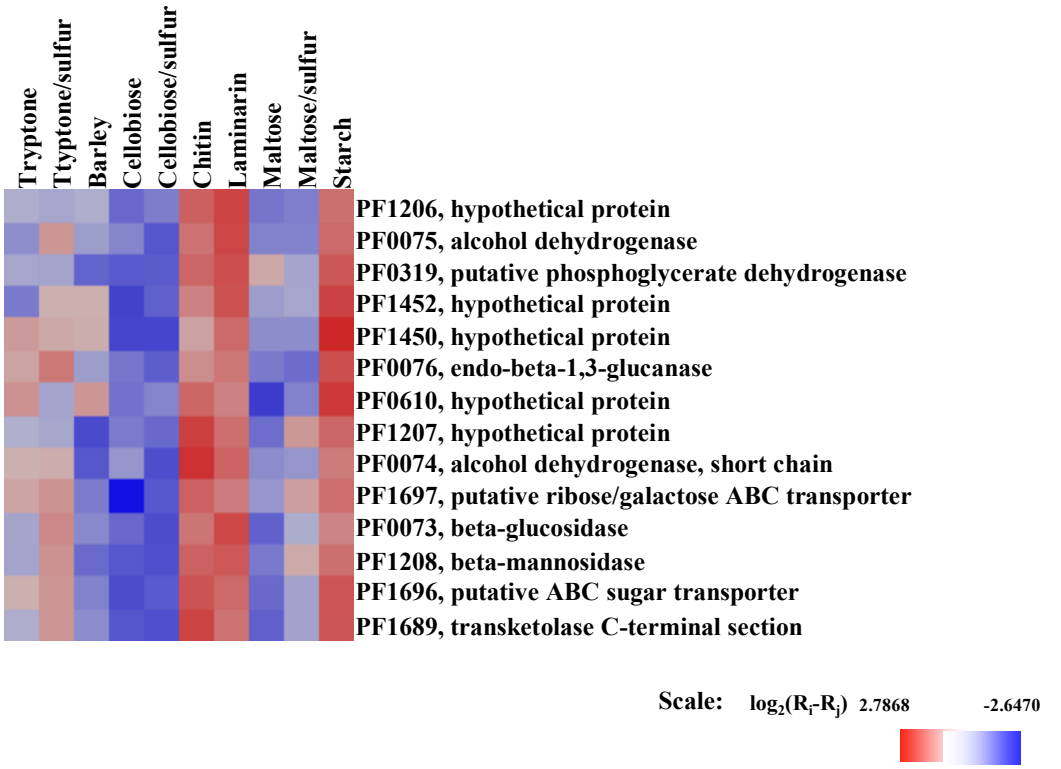


Figure 7. Cellobiose and laminarin dependent regulation. Sample clusters are constructed with standardized least square means estimates of treatment effects. Known or putative functions are indicated.

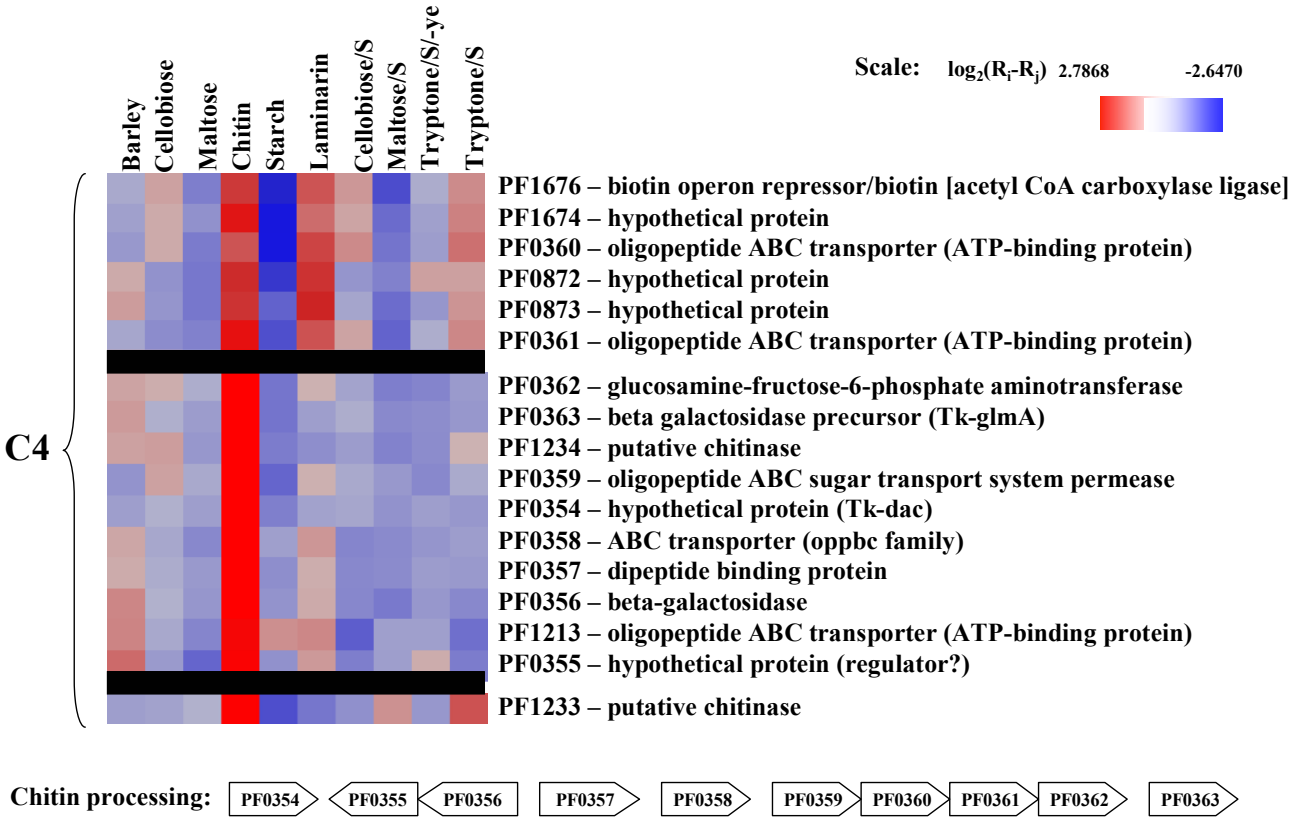


Figure 8. Chitin dependent regulation. Sample clusters are constructed with standardized least square means estimates of treatment effects. Known or putative functions are indicated. Also shown is the gene cluster containing the arrangement of open reading frames for the utilization of chitin.

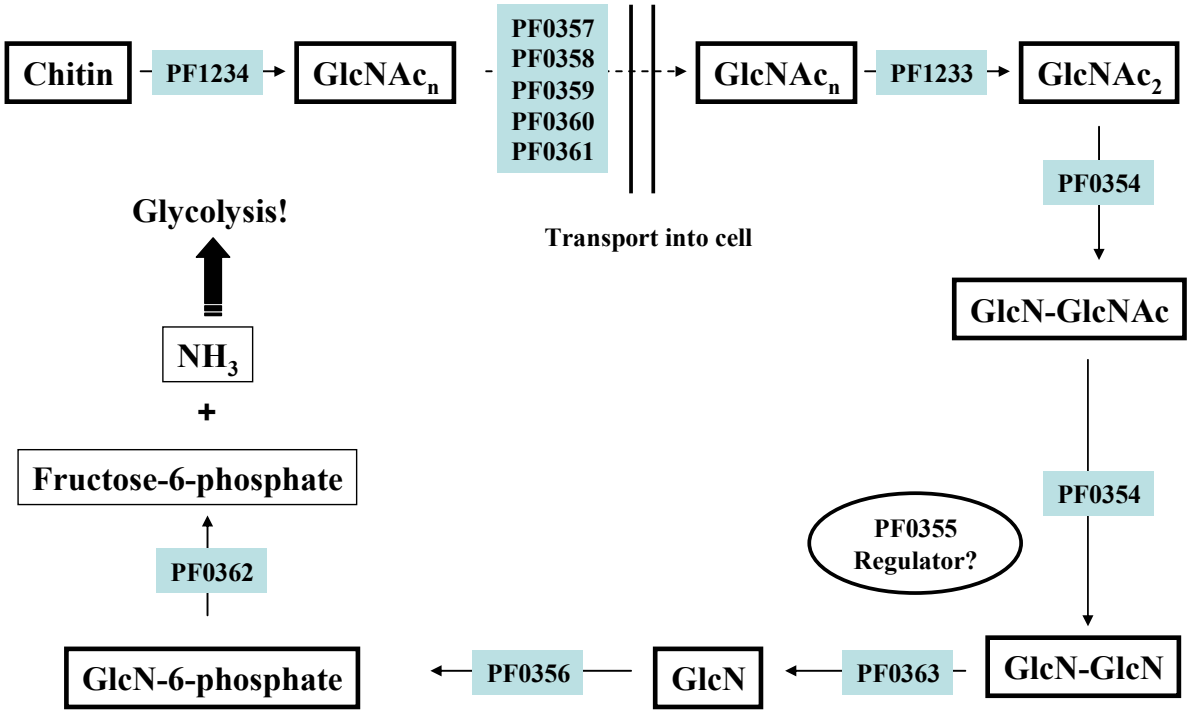


Figure 9. Proposed pathway for chitin utilization in *P. furiosus*. PF1234 encodes an extracellular, endo-acting chitinase; PF0357-PF0361 encode an ABC transporter system previously annotated as an oligopeptide transporter; PF1233 encodes an intracellular exo-acting chitinase; PF0354 encodes an glucosamine deacetylase; PF0363 encodes a glucosaminidase; PF0356 may encode a glucosamine-6-phosphate transferase which is annotated as a b-galactosidase in the genome sequence; and PF0362 encodes a glucosamine-6-fructose phosphate aminotransferase.

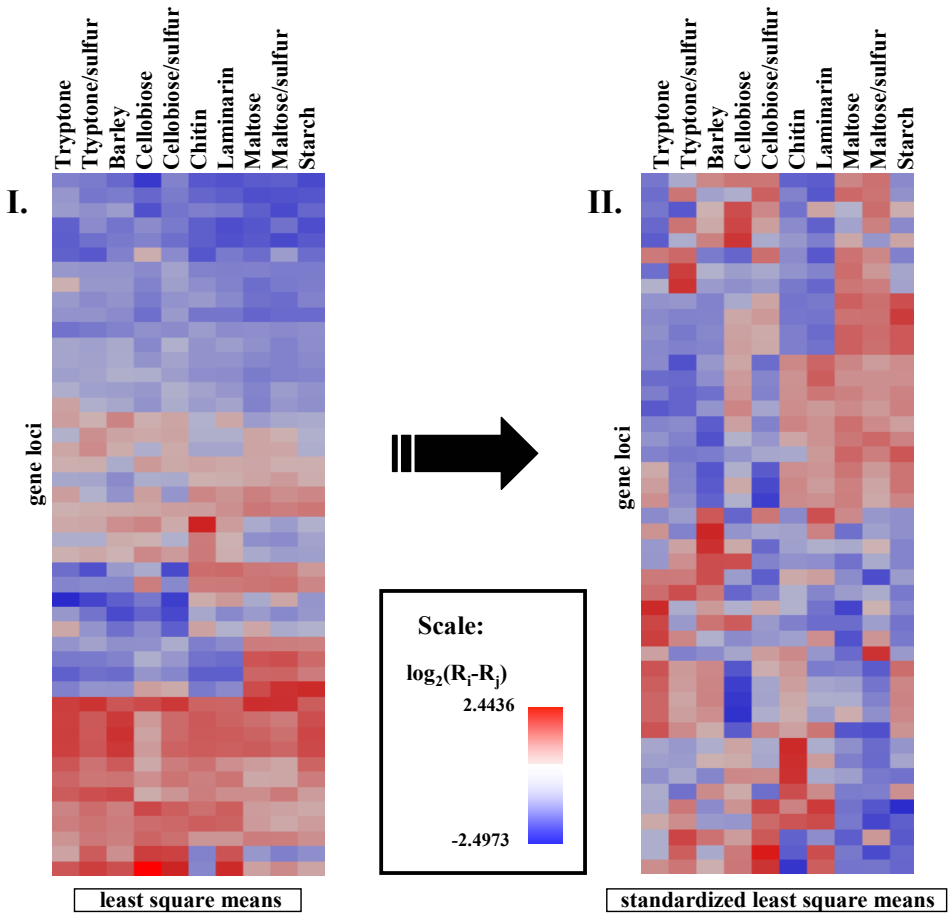


Figure 10. Regulation of ATP binding cassette transporters. The clusters represent least square means (I) and standardized least square means (II) of every annotated Opp and sugar transporter in the genome of *P. furiosus*.