

FOSSIL VIRUSES, REDOX PARADIGMS AND PREDICTIVE METABOLISM
FROM A SYSTEMS BIOLOGY PERSPECTIVE

by

Joshua Vance Heinemann

A dissertation submitted in partial fulfillment
of the requirements for the degree

of

Doctor of Philosophy

in

Biochemistry

MONTANA STATE UNIVERSITY
Bozeman, Montana

April 2014

©COPYRIGHT

by

Joshua Vance Heinemann

2014

All Rights Reserved

DEDICATION

This work is dedicated to all.

ACKNOWLEDGEMENTS

I would like to thank all the members of my committee for sharing their time, knowledge and encouragement. First I would like to thank my advisor Dr. Brian Bothner, whose support, guidance and patience not only helped me become a better scientist, but also a better person; Dr. Edward Dratz, whose perspective and guidance had a tremendous effect upon my scientific development and interests; Drs. C. Martin Lawrence and John W. Peters I thank for helping me promote my research and their guidance in thermophilic biology.

I would like to thank my collaborators for their contributions to this work, especially, Dr. Aurélien Mazurie for countless hours of data processing; Drs. Walid Maaty, Benjamin D. Reeves and Joey Steffens for their assistance with proteomics chemical labeling; Dr. Jonathan K. Hilmer and Navid Movahed for mass spectrometry training; Dr. George Gauss for assistance with cloning; Susan K. Brumfield for her help with the TEM; Dr. Mohammad J. Mohigmi for his experience in the microfabrication facility; and Brigit Noon for her tireless motivation and help on the real-time metabolomics work. A special thanks to my wife, Lindsey, and son Gabriel for their love, patience and support. This work would not have been possible without all of you.

TABLE OF CONTENTS

1. INTRODUCTION	1
Systems Biology	1
Molecular Complexity	2
Structural Complexity	2
Temporal Complexity	3
Abstraction and Modeling.....	3
Omics	5
Genomics	6
Transcriptomics/Proteomics	6
Metabolomics.....	7
Sulfolobus solfataricus as a Systems Model	8
Potential Of Omics.....	9
Mass Spectrometry.....	10
Microfluidics	10
Machine Learning	11
Robotics And Integration.....	11
Research Goals.....	12
References.....	15
2. FOSSIL RECORD OF AN HK97-LIKE PROVIRUS	20
Contribution of Authors and Co-authors	20
Manuscript Information Page	22
Abstract	23
Introduction.....	24
Results.....	27
Infection of Sulfolobus solfataricus P2 and Particle Discovery	27
Characterization of a Provirus in Sulfolobus	31
Discussion	35
Materials and Methods.....	37
Culturing of Sulfolobus solfataricus	37
Purification of Spherical Particles	37
Cloning.....	38
Expression and Purification	39
Electron Microscopy	40
Proteolysis.....	40
Identification of Putative Proviruses.....	41
Prediction, Modeling and Verification of Pseudoatomic Models.....	41
References.....	42

TABLE OF CONTENTS – CONTINUED

3. EXPANDING THE PARADIGM OF THIOL REDOX IN THE THERMOPHILIC ROOT OF LIFE	48
Contribution of Authors and Co-authors	48
Manuscript Information Page	50
Abstract	52
Introduction.....	53
Materials and Methods.....	55
Growth of Microorganisms.....	55
Zdye Maleimide Probe for thiol Labeling	56
Metabolite Extraction.....	58
LCMS based Metabolome Analysis	59
GSH Labeling and GSSG Reduction.....	60
Determination of NADPH Ratio.....	60
Results.....	61
The <i>S. solfataricus</i> Proteome is Oxidized	61
The Metabolome of <i>S. solfataricus</i> Favors Oxidized Thiol.....	62
Detection of Small Molecules Using Chemical Biology.....	64
NADPH/NADP ⁺ ratio	66
Genomic Justification for GSSG	66
Discussion	69
Acknowledgements.....	72
References.....	73
4. APPLICATION OF SUPPORT VECTOR MACHINES TO UNTARGETED METABOLOMICS EXPERIMENTS WITH LIMITED REPLICATES	77
Contribution of Authors and Co-authors	77
Manuscript Information Page	78
Abstract	79
Introduction.....	80
Materials and Methods.....	81
Culturing of <i>Sulfolobus solfataricus</i>	82
Metabolite Extraction from <i>S. solfataricus</i>	82
<i>S. scrofa</i> Urine Collection and Hemorrhagic Shock Protocol	83
Metabolite Purification from <i>S. scrofa</i> Urine	83
Untargeted LCMS Analysis.....	84
Principal Components Analysis and Hierarchical Clustering.....	85
Evaluation of Variation.....	85
SVM-RFE Data Processing	85
SVM-RFE Model Validation.....	87

TABLE OF CONTENTS – CONTINUED

Results and Discussion	87
Concluding remarks	94
Acknowledgements	95
References	96
 5. REAL-TIME DIGITIZATION OF METABOLOMIC PATTERNS FROM A LIVING SYSTEM USING MASS SPECTROMETRY	 98
Contribution of Authors and Co-authors	98
Manuscript Information Page	99
Abstract	100
Introduction	101
Materials and Methods	103
Metabolite Extraction Chip	103
Metabolite Samples	104
Mass Spectrometry	104
Interfacing the Metabolite Extraction Chip and Mass Spectrometer	105
Data Analysis	106
Results and Discussion	106
Concluding Remarks	115
Acknowledgements	116
References	117
 6. CONCLUDING REMARKS	 120
 REFERENCES CITED	 123
 APPENDICES	 132
APPENDIX A: Supporting Information for Chapter 2	134
APPENDIX B: Supporting Information for Chapter 3	139
APPENDIX C: Supporting Information for Chapter 5	144

LIST OF TABLES

Table	Page
2.1 SsP2 Protein Matches to Viral and Proviral Proteins	23
3.1 Summary of <i>S. solfataricus</i> Proteins with Similarity to Glutathione synthesis and metabolism	68
4.1 Top Ten Features as Determined by SVM-RFE	92
4.2 Training Accuracy (%) of Feature Subsets When Reducing the Number of Biological Replicates	93

LIST OF FIGURES

Figure	Page
2.1 Treatment of <i>S. solfataricus</i> with STIV Leads to a Down Regulation of Translation of Ss-vlp.....	28
2.2 Electron Micrographs of Recombinant SsP2-vlps After Negative Staining with Uranyl Acetate	30
2.3 Circular Representation of SsP2 Genome with Location of Provirus.....	34
3.1 Cysteine Labeling of Cytoplasmic Proteins of <i>S. solfataricus</i> Before and After Reduction of Disulfide Bonds.....	62
3.2 Metabolic Profiling and Characterization of Thiol Content	63
3.3 Results of Genomic, Proteomic and Metabolomic Investigation Comparing Thiol Content	70
4.1 PCA Plots and Hierarchical Clustering.....	88
4.2 Coefficient of Variation (CV) Plots for Metabolites	89
4.3 Effect of Missing Replicates on Feature Rank and Accuracy	91
5.1 Diffusion of Small Molecules on Metabolite Extraction Chip (MEC)	107
5.2 Binary Setup of Integrated Syringe Pump, Valve, Solvents, MEC, and Mass Spectrometer	108
5.3 Diffusion of GSSG (613.1 m/z) in MEC on ESI-Trap	109
5.4 Change in Metabolite Abundance Over 6 hrs. of <i>E. coli</i> Growth, for Eight Small Molecules	111
5.5 Real-Time Analysis of <i>E. coli</i> growth by MS.....	112

LIST OF FIGURES – CONTINUED

Figure	Page
5.6 Change in Small Molecule Abundance During E. coli Growth for Ten Small Molecules in three growth replicates.	113

GLOSSARY

CCMV: Cowpea Chlorotic Mottled Virus
CoA: Coenzyme-A
COG: Cluster of Orthologous Groups
CV: Coefficient of Variation
dsDNA: Double Stranded DNA
DTT: Dithioreitol
ESCRT: Endosomal Sorting Complexes Required for Transport
ESI-MS: Electrospray Ionization Mass Spectrometry
FBA: Flux Balance Analysis
GSH: Reduced Glutathione
GSSG: Oxidized Glutathione
hc: Hierarchical Clustering
HILIC: Hydrophilic Interaction Column
HSV-1: Human Herpes Simplex Virus 1 (HSV-1)
IEF: Isoelectric Focusing
LC: Liquid Chromatography
LCMS: Liquid Chromatography Mass Spectrometry
LOC: Lab-on-a-Chip
MCP: Major Capsid Protein
MEC: Metabolite Extraction Chip
MEMS: Microelectromechanical Systems
ML: Machine Learning
MRM: Multiple Reaction Monitoring
MS: Mass Spectrometry
MS-LOC: Mass Spectrometry Integrated Lab-on-a-Chip
NADP⁺: Nicotinamide Adenine Dinucleotide Phosphate (oxidized)
NADPH: Nicotinamide Adenine Dinucleotide Phosphate (reduced)
NCBI: National Center for Biotechnology Information
OD: Optical Density
ORF: Open Reading Frame
PCA: Principle Component Analysis
PDMS: Polydimethylsiloxane
Pfam: Protein Family
Pf-vlp: Virus-Like Particle from *Pyrococcus furiosus*
PSH: Protein Thiol
PSSP: Protein Disulfide Bond
Q-TOF: Quadrupole Time Of Flight
rSs-vlp: Recombinant Virus-like Particle from *Sulfolobus solfataricus*
SEC: Size Exclusion Chromatography
Ss-vlp: Virus-Like Particle from *Sulfolobus solfataricus*
SsP2: *Sulfolobus solfataricus*
SsPDO: *Sulfolobus solfataricus* Protein Disulfide Oxidoreductase

GLOSSARY - CONTINUED

SsP2-pro: *Sulfolobus solfataricus* Provirus

STIV: *Sulfolobus* Turreted Icosahedral Virus

SVM: Support Vector Machine

SVM-RFE: Support Vector Machine Recursive Feature Elimination

UHD: Ultra High Definition

UPLC: Ultra Performance Liquid Chromatography

ABSTRACT

One of the goals of systems biology is to develop a model which encapsulates the molecular, structural and temporal complexity of a living organism. While modern omics experiments can deliver a high resolution view of an organism's molecular complexity, methods for correlating the information from multiple biomolecular systems (i.e. genes, proteins and metabolites) and their changes over time remain greatly underdeveloped. Presented in this research are: (1) methods for understanding the inter-relation of multiple biomolecular systems correlating genomics, proteomics and metabolomics experiments; (2) techniques for machine learning based metabolic biomarker selection; (3) robotics technology for real-time measurement of changes in metabolism. The methods for correlating information from multiple biomolecular systems have provided a new perspective of biomolecular adaptation and evolutionary relationships in the thermophilic archaea. The techniques for biomarker selection have provided a method to assess the reliability of biomarkers in experiments where limited samples are available. The new technology has provided an engineered system for automated analysis of metabolic patterns and how they change over time. Together, these results have created a framework for future improvement of our understanding of biology through the use of molecular biology, machine learning and robotics.

CHAPTER 1

INTRODUCTION

Systems Biology

Systems biology is an emerging paradigmatic design adopted by the scientific community for the study of biomedical and biological scientific research. This design dictates transcendence from any one scientific discipline to encompass a holistic interdisciplinary perspective of biology. Using this interdisciplinary perspective we shift from the reductionist approach of looking at single components (genes, proteins, metabolites) to an all-inclusive model for understanding living systems. The adoption of systems biology has largely been driven by technological innovations, particularly in genome sequencing and high-throughput abundance measurements using mass spectrometry (MS). Aided by this technology, mathematical and abstract models that can predict the behaviors of life through the integration of information from multiple biomolecular systems have been constructed. That being said, the dimensions of systems biology are vast and even as we assemble more detailed knowledge of the parts, we cannot simply linearly add them up to model a living organism. However, it is helpful to break down the dimensions of systems biology into categories we can physically encompass, to help rationalize behaviors and gain insight to the emergent properties that manifest in everyday life. Here I suggest we can describe systems biology with three

separate categories: (1) molecular complexity, (2) structural complexity, and (3) temporal complexity; similar to those suggested by (Huang et. al., 2006).

Molecular Complexity

Molecular complexity refers to the complexity which comes from the cell as the sum of all of its molecular parts i.e. DNA, RNA, proteins, and metabolites. As such, one of the major goals of research in this area is to establish a piece by piece comprehensive network map of all regulatory interactions between these parts (Davidson et al., 2002). The deduction of these complex networks is much easier said than done, as assembly can only truly be completed with the arduous demonstration of individual interactions through experimentation. Many networks have been built for generation of theoretical models using genome based homology, but insights gained from these remain limited because of a lack of in vivo experimental data (Barabasi and Albert et. al., 2009, Barabasi et. al., 2004). It becomes obvious when looking at the genomes of model organisms that functions of a living system are determined not only by individual genes and their encoded proteins, but also by how evolution has shaped the interactions at the molecular level.

Structural Complexity

While comprehensive omics study of complex systems emphasizes the network architecture and molecular processing aspect of interacting biomolecules, the complexity of an organism must also take into account the multiple size scales across a wide range of space (Bar-Yam et. al., 2004). Metabolites being the smallest (Da), are used to build

genes and proteins (kDa), who assemble into complex structures (MDa), which organize into cells (μm). Cells assemble into biofilms or tissues (mm) which can make up multicellular organisms (m). The orientation and location in each of these three dimensional structures (from nm to m) allows precise control of the organism as a whole. This aspect of a living system is invisible to genomic based science and is traditionally considered the domain of biophysics.

Temporal Complexity

Time is probably the least understood dimension in systems biology. The properties of both molecular and structural complexity vary with time, the dimension of dynamic behavior. Physical and metabolic processes exhibit characteristic temporal changes of varying time scales. Electron transfer happens on the femtosecond timescale (PHz) (Beratan et. al., 1991), protein conformational changes nanoseconds (GHz), ion channel gating events occur in microseconds (MHz), the cardiac cycle every second (Hz), and all can directly affect the physiology of an organism which can live on the order of gigaseconds. The timescale depth of even one organism confounds our current ability to accurately model each individual interaction, a depth that does not even begin to encompass how time has influenced evolution over the last 3.7 billion years (Ohtomo et. al., 2013).

Abstraction and Modeling

Because of their complexity, biological dimensions have been modeled either with a reductionist approach where small chunks of the cell are modeled with the utmost

detail (Barik et. al., 2010, Chen et. al., 2004, Kuttykrishnan et. al., 2010), or with abstract models using simplified representations to describe the emergent properties of living systems (Tyson et al. 2003, Voit et al. 1991, Savageau et al. 2009, Soni et al. 2008, Margolin et al. 2006).

For reductionist models, flux balance analysis (FBA) is one of the more popular mathematical approaches to analyzing the flow of metabolites throughout a system, using mathematical formalism as a shortcut (Orth et al. 2010). Even in their reduced complexity, these models have limitations as several common scenarios cannot be accounted for. Such as, the intracellular pathway is treated as unidirectional (Feng et al. 2012), parallel metabolic routes and cyclic fluxes cannot be resolved, and futile cycles cannot be represented in the solution (Steuer et al. 2009).

In contrast, abstract models try to explain emergent system features or observed collective behavior that cannot be explained by studying the dissected parts. Examples of emergent system behaviors include the collective oscillations observed in yeast (Richard et al. 2003, Chen et al. 2004, Barik et al. 2010) and the maintenance of intracellular redox homeostasis (Sevier et al. 2002). These behaviors are of particular interest because they inevitably lead to predictable spatiotemporal patterns. How these patterns emerge is often unknown, but they must be created by the particular behaviors and interactions of components at a lower level.

Omics

The adoption of systems biology has probably been driven by the development of omics technology. These datasets have greatly expanded our ability to understand biological complexity at the molecular, structural and temporal levels. Starting with genomics which saw the sequencing of the first genome, bacteriophage ϕ X174 in 1977 (Sanger et al. 1977), transcriptomics gained popularity, then proteomics and more recently metabolomics. Omics data, which includes genomics, transcriptomics, proteomics and metabolomics provides a bird's eye view of the low level molecular complexity in the respective biomolecular system. However, only abstract models can be obtained from this data, as detailed reaction kinetics and spatiotemporal changes typically are not quantified. Thus the ability to generate a mathematically predictive model using omics data is currently not possible, as only the existence of biomolecules and their abundance patterns can be observed. Despite this, omics based abstract models can offer valuable biological insight. In adaptive biological systems, stress or disease can cause expression changes of RNA, proteins and metabolites which can be detected (Ortmann et al. 2008, Maaty et al. 2009, and Patti et al. 2012). These low level regulatory changes provide the biomolecular identities of those responsible for maintenance of homeostasis in the presence of the disease or stress. Thus, each of the omic categories can provide an important biological perspective.

Genomics

The first DNA sequences were obtained in the 1970s using two-dimensional chromatography (Sanger et. al., 1977), but the technology has evolved to allow next-generation methods which are automated and orders of magnitude faster. The demand for fast and low-cost genome sequences has made 2014 the year of the thousand dollar human genome (Hayden et. al., 2014) only forty-four years after the first full genome sequence (Sanger et. al., 1977). Genomic sequencing has revolutionized our perspective of biology from multiple angles. The rapid speed of DNA sequencing attained with modern technology has been instrumental in the mapping the genomes of thousands of organisms. We now know the range of genome based molecular complexity for sequenced organisms, the smallest currently being *Candidatus Carsonella ruddii* (<160 kbp) (Nakabachi et al. 2006) and largest *Paris japonica* (>150 Gbp) (Pellicer et al. 2010). Genomics has revolutionized microbial taxonomy, and lead to the identification of a third domain of life, the Archaea (Woese et al. 2000). Additionally, the large number of genome sequences gives us a radical new view of molecular evolution and information how DNA is shared across species boundaries (Heinemann et. al., 2011).

Transcriptomics/Proteomics

The transcriptome is the set and concentration of all RNA molecules, including the mRNA, rRNA, tRNA, and other non-coding RNA produced from the genome of a given population of cells. Initially the transcriptome was surveyed using DNA microarrays (Augenlicht et al. 1982, Schena et al. 1995), but like genomics has adapted to next-generation sequencing techniques to allow more comprehensive survey (Chu et al.

2012). The transcriptome can be seen as the precursor to the proteome, the entire set of proteins expressed. Expression profiling has become popular with both transcriptomics and proteomics, where the measurement of the expression of thousands of genes (or proteins) can be done simultaneously. These profiles, for example, can distinguish between cells which have been stressed with hydrogen peroxide (Maaty et al. 2009) or infected with virus (Ortmann et al., 2008, Maaty et al., 2012), and determine the metabolic pathways involved. While the correlation of transcriptomics and proteomics seems like it should create redundancy, many differences such as alternative splicing, posttranslational modification and a difference in temporal timescales can create a disparity between the two expression systems (Gedeon et al. 2012).

Metabolomics

Of the omics categories described, metabolomics is the newest to gain popularity and arguably the hardest to correlate to the other omics levels (Fiehn et al. 2002). This is because tens of thousands of unknown small molecule metabolites can be detected and unlike genomics, transcriptomics and proteomics which have a common unifying code, small molecule metabolites are not so easily mapped to their interacting partners in other systems (Yanes et al. 2010, Want et al. 2006). Metabolic abundances also appear more dynamic in time, although comprehensive high time resolution data for the other omics data is also somewhat lacking. Regardless of their correlation, the scope of this data makes for no easy interpretation.

Sulfolobus solfataricus P2
as an Omic Systems Model

S. solfataricus has emerged as one system of choice for correlating multi-omics datasets in the thermophilic archaea for several practical reasons. It is an aerobic heterotroph that can be easily cultured to high cell densities in either batch or continuous culture, with a reported doubling time of approximately four hours (Grogan et al. 1989). The genome of the organism has been sequenced (She et al. 2001), and a rudimentary network model of its central carbon metabolism has been assembled (Snijders et al. 2006). This has led to *S. solfataricus* use as an archaeal systems model for studying viral infection (Ortmann et al. 2008, Maaty et al. 2012), and oxidative stress (Maaty et al., 2009).

The era of omics technology has generated a lot of new biological insight on many organisms, but comparisons in archaea like *S. solfataricus* provides a unique perspective on the emergent and divergent properties in the third domain of life. Using integrated omic technology a general coordination of results between the expression changes in transcriptomics and proteomics could be observed (Ortmann et al. 2008, Maaty et al. 2009, Maaty et al. 2012). Expression profiling during viral infection has shown a coordinated response in up regulation of the endosomal sorting complexes required for transport (ESCRT), a property also observed in eukaryotic viral infection (Maaty et al. 2012). Coordinated transcriptomic and proteomic expression changes were also seen in the oxidative stress response, validating a novel regulatory system in this organism. Thus, coordinated omics data between transcriptomic and proteomic systems

has already shown the adaptation of biomolecular machinery to thermophilic life which is both similar and vastly different than those observed in other domains of life.

Potential of MS-based Omics

The development of omics technology has indisputably been the driving force behind its popularity, as accessibility and ease of use has facilitated its adaptation in many research laboratories worldwide. The omics described provides a high resolution view of the molecular complexity which exists in multiple cellular systems. These technologies are much less informative in answering questions about structural or temporal complexity. While structural complexity is probably better studied with directed fluorescent labeling within living cells, classical omics techniques using mass spectrometry are suitable for study of the relationship between molecular and temporal complexity. However, to view temporal complexity a different strategy must be adopted in the development of these technologies. MS-based omics aim at being comprehensive in their detection of biomolecules from the respective system. This is good for viewing molecular complexity, but bad for viewing temporal complexity as time consuming separations are required on top of sample preparation. The inhibitory speed of preparation and analysis creates difficulty in viewing higher time resolution. Modern technology provides the capability of overcoming these shortfalls and stands to revolutionize biological study through the integration of several important functions.

Mass Spectrometry

Mass spectrometry is an analytical technique that produces spectra of the masses of atoms or molecules in a sample or material. Spectra can be used to determine the identity, abundance, isotopic signature and chemical structure of biomolecules on millisecond timescales (MHz). MS does not require the arduous development of chemical sensors as it can be used to detect anything of the appropriate size that ionizes, which includes many biomolecules. Because of the wide-scale use of MS in proteomics and metabolomics it has been optimized for the detection of biomolecules and generation of large biological datasets. Additionally, MS has also been engineered using microelectromechanical systems (MEMS) technology so that it can exist on a miniaturized chip, with recent developments steadily increasing mass range and resolution (Syms et al. 2009).

Microfluidics

Microfluidic or lab-on-a-chip (LOC) technology provides an interface between MS and biological systems (Gao et al. 2013). LOC functions have been demonstrated to manipulate cellular systems in many ways not possible with other technologies, with the added benefit of being microscale (Macounova et al. 2000, Bhagat et al. 2010). These systems can be engineered to be reproducible and biocompatible enabling long term separation and manipulation of biomolecules (Bhagat et al. 2010). Microfluidics has been instrumental in the development of some next-generation sequencing machines as its tiny size allows for massive parallelization to be built in a small space (Voelkerding et

al. 2009). Additionally LOC valving has been demonstrated to operate on millisecond timescales (MHz) making it easily adaptable to MS (Melin et al. 2007).

Machine Learning

Machine learning (ML) is a branch of artificial intelligence that focuses on the learning of computers from data. While ML is a broad category that encompasses many forms of computational learning, it has been particularly effective in complex pattern recognition and generalization. Generalization is the property that the system will perform well on unseen data. Support vector machines (SVM) have become a popular form of ML, which analyze data and recognize patterns used for classification and regression analysis (Cortes et al. 1995). Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns unseen examples into the most appropriate category. In the context of omics datasets an SVM can be trained to identify the patterns associated with any given physiological condition (Guyon et al. 2002). They have the added bonus of being applicable to datasets where the number of training examples is small and the variables associated with the examples is large ($n \gg p$), datasets commonly observed in omics.

Robotics and Integration

A robot is typically an electro-mechanical machine that is guided by a computer program or electronic circuitry to accomplish a task. Robotic systems can be designed to take the place of humans in repetitive tasks. Robot functionality has already been integrated in with MS through the computational control of robotic arms located in

autosamplers. This robot allows a user to set a large number of samples into a holding chamber until the appropriate time when the robot arm will inject the sample for MS analysis. This semi-autonomous robot has greatly increased the number of samples which can be analyzed in MS-based metabolomics as the robot can be programmed to operate twenty-four hours a day, seven days a week without stopping. This robotic functionality has streamlined the data collection process, and carefully engineered functions for fast sample preparation have the potential to further increase both the reproducibility and time resolution of MS based omics.

Robotic integration of MS, microfluidics and machine learning has the potential to automate biological study using a carefully engineered system (Soldatova et al. 2006, King et al. 2009). This would be ideal for the study of metabolism as evidence suggests a large degree of natural metabolic flexibility exists in an organism.

Research Goals

The work presented in this dissertation was directed with two major goals. The first goal was to provide biological insight into the evolution of the model organism *S. solfataricus* and its adaptation to life at extreme temperatures using the integration of information from multiple biomolecular systems. Previous work has provided insight into how changes in molecular complexity allow the organism to cope with viral infection (Ortmann et al. 2008, Maaty et al. 2012) by expression profiling at both transcriptomic and proteomic levels. This research has identified biomolecular targets which share homology with metabolic pathways from other domains of life, and others

which appear to be divergent systems. However, large gaps in our understanding remain, such as, how previous viral infections have affected the evolution of this organism. The genomic landscape of many organisms contains information about the evolutionary effect of viral mediated horizontal gene transfer, but how far reaching it is in thermophilic archaea? In addition, expression profiling under oxidative stress (Wiedenheft et al. 2005, Maaty et al. 2009) suggested that *S. solfataricus* uses a novel system for coping with oxidative stress, but the explanation which justifies the need for this system was not well understood. This organism lives in an environment which would be fatal for most life, therefore how has the defense from oxygen free radicals affected *S. solfataricus* survival strategy, if at all.

The second goal was to develop meaningful insight about how metabolic patterns change in time and in response to stress. This goal can be separated into two different questions. The first question is can machine learning based pattern recognition, which is typically used for facial recognition; find conserved patterns and biomarkers in metabolomic datasets? This is a complex problem because machine learning algorithms perform best with a large number of training examples and few variables. However, most biological experiments (to include metabolomics) have only a limited number of replicates and a large number of variables (metabolites).

The second question is how much metabolic variability is intrinsic to an organism? The answer will allow us to determine if a biomarker can be a reliable indicator in every possible metabolic configuration. Previous evidence suggested that organisms have natural metabolic variability which can be encapsulated into a theoretical

metabolic phase space (Schmidt et al. 2011). The theory of phase space was proposed because there are natural oscillations in metabolite abundance. Best studied in yeast, metabolic oscillations have been seen in bacteria (Levering et al. 2013), plants (Farre et al. 2005, Henson et al. 1990), insects (Schwemmler et al. 1980), mice (Dahlgren et al. 2005) and humans (Bergesten et al. 2002, Iotti et al. 2010, Berk et al. 2011). Natural metabolic oscillations have not been studied on a large scale, mostly because of a lack of technology for tracking these changes over time.

References

- Augenlicht, Wahrman MZ, Halsey H, Anderson L, Taylor J, Lipkin M, (1987) Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro, *Cancer Research*. 47, 6017–21.
- Barabasi AL, Albert R, (1999) Emergence of scaling in random networks, *Science*. 286, 509–512
- Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization, *Nature reviews genetics*. 5, 101–113.
- Barik D, Baumann W T, Paul M R, Novak B, Tyson JJ (2010) A model of yeast cell-cycle regulation based on multisite phosphorylation, *Molecular systems biology*. 6, 405.
- Bar-Yam Y, (2004) Multiscale variety in complex systems. *Complexity*. 9, 37–45.
- Beratan DN, Betts JN, and Onuchic JN, (1991) Protein electron transfer rates set by the bridging secondary and tertiary structure. *Science* 252, 1285-88.
- Bergesten P, (2002) Role of oscillations in membrane potential, cytoplasmic Ca²⁺, and metabolism for plasma insulin oscillations, *Diabetes*. 51, Suppl 1: S171-6.
- Berk M, Ebbels T, Montana G, (2011) A statistical framework for biomarker discovery in metabolomics time course data, *Bioinformatics*. 27(14), 1979-85.
- Bhagat AS, Bow H, Wei Hou H, Jin Tan S, Han J, Lim C (2010) Microfluidics for cell separation, *Medical & biological engineering and computing*. 48, 999-1014.
- Chen K C, Calzone L, Csikasz-Nagy A, Cross F R, Novak B and Tyson J J (2004) Integrative analysis of cell cycle control in budding yeast, *Molecular biology of the Cell*. 15, 3841–62
- Chu Y, Corey DR (2012). RNA sequencing: platform selection, experimental design, and data interpretation, *Nucleic acid therapeutics*. 22, 271-4.
- Cortes C, Vapnick V, (1995) Support-vector networks, *Machine learning*. 20, 273-97.
- Dahlgren GM, Kauri LM, Kennedy RT, (2005) Substrate effects on oscillations in metabolism, calcium, and secretion in single mouse islets of Langerhans, *Biochimica et biophysica acta*. 1724(1-2), 23-26.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R, Rust AG, Pan ZJ, Schilstra MJ, Clarke PJC, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H (2002) A genomic regulatory network for development, *Science*. 295, 1669–78.

Farre EM, Weise SE, (2012) The interactions between the circadian clock and primary metabolism, *Current opinion in plant biology*. 15(3), 293-300.

Feng X, Xu Y, Chen Y, Tang YJ, (2012) Integrating flux balance analysis in kinetic models to decipher the dynamic metabolism of *Shewanella oneidensis* MR-1, *PLoS Computational Biology*. 8, e1002376.

Fiehn O (2002). Metabolomics - the link between genotypes and phenotypes. *Plant molecular physiology*. 48 155-171.

Gao D, Liu H, Jiang Y, Lin J (2013) Recent advances in microfluidics combined with mass spectrometry: technologies and applications, *Lab on a chip*. 13, 3309-22.

Gedeon T, Bokes P (2012). Delayed protein synthesis reduces the correlation between mRNA and protein fluctuations, *Biophysical journal*. 3, 377-85.

Grogan, D. W. (1989) Phenotypic Characterization of the Archaeobacterial Genus *Sulfolobus* - Comparison of 5 Wild-Type Strains, *Journal of Bacteriology*. 171, 6710-6719.

Guyon I, Weston J, Barnhill S, (2002) Gene Selection for cancer classification using support vector machines, *Machine learning*. 46, 389-422.

Hayden EC, (2014) Is the \$1000 genome for real? *Nature*. doi:10.1038

Henson CA, Duke SH, (1990) Oscillations in plant metabolism, *Progress in clinical and biological research*, 341B, 821-34.

Huang, S., Wikswo, J., (2006) Dimensions of systems biology. *Reviews of physiology biochemistry and pharmacology* 157, 81-104.

Iotti S, Borsari M, Bendahan D, (2010) Oscillations in energy metabolism, *Biochimica et biophysica acta*. 1797(8), 1353-61.

King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, Sparkes A, Whelan KE, Clare A, (2009) The automation of science, *Science*. 324, 85-9.

Kitano, H., (2002) Systems biology: a brief overview. *Science* 295, 1662-1664.

Kuttykrishnan S, Sabina J, Langton LL, Johnston M and Brent MR, (2010) A quantitative model of glucose signaling in yeast reveals an incoherent feed forward loop leading to a specific, transient pulse of transcription, *Proceeding of the national academy of science of the United States of America*. 107, 16743–8.

Levering J, Kummer U, Becker K, Sahle S, (2013) Glycolytic oscillations in a model of lactic acid bacterium metabolism, *Biophysical Chemistry*. 172, 53-60.

Maaty WS, Heinemann J, Bothner B, (2009) Something old, something new, something borrowed; how the thermoacidophilic archaeon *Sulfolobus solfataricus* responds to oxidative stress, *PloS one*. 4, e6964.

Maaty WM, Selvig K, Ryder S, Tarlykov P, Hilmer JK, Heinemann J, Steffens J, Snyder JC, Ortmann AC, Movahed N, Spicka K, Chetia L, Grieco PA, Dratz EA, Douglas T, Young MJ, Bothner B. (2012) Proteomic analysis of *Sulfolobus solfataricus* during *sulfolobus* turreted icosahedral infection, *Journal of proteome research*. 11, 1420-32.

Macounova K, Cabrera CR, Holl MR, Yager P, (2000) Generation of natural pH gradients in microfluidic channels for use in isoelectric focusing, *Analytical chemistry*. 72, 3745-51.

Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, la Favera R and Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinformatics*. 7, Suppl. 1 S7.

Melin J, and Quake SR, (2007) Microfluidic large-scale integration: the evolution of design rules for biological automation, *Annual review of biophysics and biomolecular structure*. 36, 213-31.

Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar H, Moran N, Hattori M (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*, *Science*. 314, 267.

Obrig H, Neufang M, Wenzel R, Kohl M, Steinbrink J, Einhaupl K, Villringer A, (2000) Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults, *Neuroimage*. 12(6), 623-39.

Ohtomo, Y., Kakegawa, T., Ishida, A., Nagase, T., Rosing, M.T., (2013) Evidence for biogenic graphite in early archaean isua metasedimentary rocks, *Nature Geoscience*. 7, 25-28.

Orth JD, Thiele I, Palsson BO, (2010) What is flux balance analysis? *Nature Biotechnology* 28, 245-248.

Ortmann, AC, Brumfield SK, Walther J, McInnerney K, Brouns SJ, Harmen, van de Werken, HJ, Bothner B, Douglas T, van de Oost J, Young MJ, Transcriptome analysis of the archaeon *Sulfolobus solfataricus* with *Sulfolobus* turreted icosahedral virus, *Journal of virology*. 82, 4874-83.

Patti GJ, Tautenhahn R, Siuzdak G, (2012) Meta-analysis of untargeted metabolomic data from multiple profiling experiments, *Nature Protocols*. 7508–16.

Pellicer J, Fay MF, Leitch IJ, (2010) The largest eukaryotic genome of them all? *Botanical journal of the Linnean Society*. 164, 10-15.
Richard P, (2003) The rhythm of yeast, *FEMS microbiology reviews*. 27, 547-57.

Savageau MA, (2009) *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. Reprint of the Original Edition Published in 1976 (Reading, MA:Addison-Wesley).

Schena M, Shalon D, Davis RW, Brown PO, (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*. 270, 467-470.

Schwemmler W, Herrman M, (1980) Oscillation in the energy metabolism of the insect host and symbiont II. Analysis of possible endogenous rhythms in both systems, *Cytobios*. 27, 193-208.

Sevier CS, Kaiser C (2002) Formation and transfer of disulphide bonds in living cells, *National review molecular cell biology*. 3, 836–47.

Soldatova LN, Clare A, Sparkes A, King RD, (2006) An ontology for a robot scientist, *Bioinformatics*. 22, e464-71.

Soni AS, Jenkins JW and Sundaram S S (2008) Determination of critical network interactions: an augmented Boolean pseudo-dynamics approach, *IET Systems Biology*. 2, 55–63.

Steuer, R., Junker, B.H., Computational models of metabolism: stability and regulation in metabolic networks, *Advances in chemical physics*. 142, ISBN-10: 0-470-46499-2.

Syms, R. R. A., (2009) Advances in microfabricated mass spectrometers, *Analytical and Bioanalytical Chemistry*. 393, 427-29.

Tyson JJ, Chen KC, Novak B, (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell *Current opinion in cell biology*. 15, 221–31.

Voelkerding KV, Dames SA, Durtschi JD, (2009) Next-generation sequencing: from basic research to diagnostics, *Clinical chemistry*. 55, 641-58.

Voit EO, (1991) *Canonical Nonlinear Modeling: S-System Approach to Understanding Complexity* (New York: VanNostrand Reinhold)

Want EJ, O'maille G, Smith CA, Brandon TR, Uritboonthai W, Qin C, Trauger SA, Siuzdak G, (2006) Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry, *Analytical chemistry*. 78, 743-52.

Weidenheft B, Mosolf J, Willits D, Yeager M, Dryden KA, Young M, Douglas T, (2005). An archaeal antioxidant: characterization of a Dps-like protein from *Sulfolobus solfataricus*, *Proceedings of the National Academy of Science of the United States of America*. 102(30), 10551-6.

Woese CR (2000). Interpreting the universal phylogenetic tree, *Proceedings of the National Academy of Sciences of the United States of America*. 97, 8392-96.

Yanes O, Tautenhahn GJ, Patti G, Siuzdak G, (2011). Expanding coverage of the metabolome for global metabolite profiling, *Analytical chemistry*. 83, 2152-61.

CHAPTER TWO

FOSSIL RECORD OF AN HK-97-LIKE PROVIRUS

Contribution of Authors and Co-Authors

Manuscript in Chapter 2

Author: Joshua Heinemann

Contributions: Virus particle purification; cloning and expression; Prediction, modeling, and verification of pseudoatomic models; bioinformatic characterization of provirus; wrote manuscript.

Co-Author: Walid S. Maaty

Contributions: Culturing of *S. solfataricus*; virus particle purification; provided valuable insight; edited manuscript.

Co-Author: George Gauss

Contributions: Assisted with cloning and expression; provided valuable insight.

Co-Author: Narahari Akkaladevi

Contribution: recombinant virus particle purification.

Co-Author: Susan K. Brumfield

Contribution: Took TEM images of native virus-like particle.

Co-Author: Vamseechar Rayaprolu

Contribution: Took TEM images of recombinant virus-like particle.

Co-Author: Mark Young

Contributions: Provided valuable insight; edited manuscript.

Contribution of Authors and Co-authors Continued

Co-Author: C. Martin Lawrence

Contributions: Provided valuable insight; edited manuscript.

Co-Author: Brian Bothner

Contributions: Provided valuable insight; wrote manuscript; procured funding

Manuscript Information Page

Authors: Joshua Heinemann, Walid S. Maaty, George Gauss, Narahari Akkaladevi, Susan K. Brumfield, Vamseedhar Rayaprolu, Mark Young, C. Martin Lawrence and Brian Bothner

Journal: Virology

Status of the manuscript:

☐ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☐ Accepted by a peer-reviewed journal

☒ Published in a peer-reviewed journal

Publisher: Elsevier

Issue in which manuscript appears: Volume 417(2), 362-8 (2011)

FOSSIL RECORD OF AN HK97-LIKE PROVIRUS

Joshua Heinemann^a, Walid S. Maaty^a, George Gauss^b, Narahari Akkaladevi^b, Susan K. Brumfield^c, Vamseedhar Rayaprolu^a, Mark Young^c, C. Martin Lawrence^b and Brian Bothner^{a*}

^{a,b,c} Department of Chemistry and Biochemistry, Montana State University, 126 Chemistry Biochemistry bldg., Bozeman, MT 59717, USA (406)994-5270

Abstract

One of the outstanding questions in biology today is the origin of viruses. We have discovered a protein in the hyperthermophile *Sulfolobus solfataricus* while following proteome regulation during viral infection that led to the discovery of a fossil provirus. Characterization of the wild type and recombinant protein revealed that it assembled into virus-like particles with a diameter of ~32 nm. Sequence and structural analysis showed that the likely proviral capsid protein, Sso2749, is homologous to a protein from *Pyrococcus furiosus* that forms virus-like particles using the HK-97 major capsid protein fold. The SsP2-provirus has a high degree of mosaicism and contains proteins with similarity to, among others, eukaryotic herpesviruses and tailed dsDNA bacteriophage families, reinforcing the hypothesis of a common ancestral gene pool across all three domains of life. This is the first description of the HK-97 fold in a crenarchaeal virus and the first direct genomic connection of linocin-like protein cages to a virus.

Introduction

It has been suggested that the viral universe can be organized into a limited number of lineages. But, with an estimated 10^{31} virus particles on earth (Suttle, 2005) and only 2443 viral genomes in the National Center for Biotechnology Information (NCBI) database, the current data is limited. Recent studies of the human genome indicate that a significant portion of our genetic information is probably of viral origin and it seems likely that viruses have strongly influenced the evolution of most, if not all, cells through the transfer of genetic elements (Brussow, Canchaya, and Hardt, 2004; Campbell, 2003; Chibani-Chennoufi et al., 2004; Dunigan, Fitzgerald, and Van Etten, 2006; Forde et al., 2008; Hatfull, 2008; Hendrix, 2003; Jiang and Paul, 1998; Lin and Li, 2009; Moreira, 2000; Rokytá et al., 2006). The growing number of sequenced genomes (viral and proviral) has enabled the construction of viral genome based phylogenies (Hatfull, 2008; Krupovic, Forterre, and Bamford, 2010). Genome level comparisons are sufficient to establish relationships for closely related viruses; however, for distantly related viruses, phylogenetic relationships must be based upon more highly conserved features, such as protein structure and function (Bamford, 2003; Bamford, Burnett, and Stuart, 2002; Krupovic, Forterre, and Bamford, 2010). Identifying these conserved features helps define evolutionary relationships between distantly related viruses, and may provide better understanding of how virus-host relations have influenced evolution.

Due to the small size of viral genomes and their ubiquity, viruses are an excellent case study for investigating genomic stability. While viruses must retain a minimal set of core functions to ensure replication, some families such as the tailed dsDNA viruses,

have highly plastic genomes. This large family appears to be both a reservoir and conduit for exchange of genetic information throughout the entire biosphere (Bamford, 2003). In depth genomic comparison of viruses from this family has provided valuable information about their mosaic nature and capacity for horizontal gene transfer in the host population (Hatfull, 2008). Amino acid sequence comparison has revealed a relationship between the tailed dsDNA viruses infecting bacteria and archaea (Krupovic, Forterre, and Bamford, 2010; Prangishvill, Garrett, and Koonin, 2006). Icosahedral virus-like particles (vlps) predicted to be of viral origin have been discovered and several examples of tailed dsDNA viruses have been observed in the Euryarchaeaota, but there have been no reports from Crenarchaeaota (Namba et al., 2005). These vlps are composed of subunits with a distinct similarity to the major capsid protein (MCP) of the bacteriophage HK97 (Akita et al., 2007). The HK97 MCP fold is found in bacterial viruses, encapsulin protein cages, and shares structural similarity to the human herpes simplex virus 1 (HSV-1) VP5 floor domain (Bamford, Grimes, and Stuart, 2005; Krupovic, Forterre, and Bamford, 2010; Sutter et al., 2008). Expanding upon the similarities of protein sequence and structure in tailed dsDNA viruses, to include DNA packaging (Rao and Feiss, 2008) and assembly (Heymann et al., 2003) has led to the proposal that similar viruses from all domains of life may have descended from a common ancestor (Bamford, 2003; Bamford, Burnett, and Stuart, 2002; Bamford, Grimes, and Stuart, 2005; Krupovic, Forterre, and Bamford, 2010). These familial ancestors are thought to be the source of the common characteristics, which are referred to as the “viral self”, and provide insight to evolutionary processes (Bamford, 2003).

The three domains of life can be split into Bacteria, Archaea, and Eucarya based on a 16S ribosomal RNA phylogenetic tree. Using this approach, Archaea are more closely related to Eucarya than Bacteria (Woese, 2000). Because virus-host systems coevolve, it seems reasonable that viruses from each of the three domains would share a similar phylogenetic relationship. Previously described classification systems have organized viruses into families based upon genome similarity arising from a shared gene pool (Hatfull, 2008). Expanding on this, we provide evidence that the tailed dsDNA virus family from bacteria has shared a gene pool with the eukaryotic herpesvirus family and archaeal viruses through a newly discovered provirus found in the Crenarchaeote *Sulfolobus solfataricus* (SsP2). The genome of the SsP2 provirus (SsP2-pro) is a mosaic of proteins with similarity to those of herpesvirus and tailed dsDNA virus families. Additionally the SsP2-pro major capsid protein (MCP) shares predicted structural similarity with MCP gp5 of the HK97 bacteriophages, encapsulin protein, the virus-like particle from *Pyrococcus furiosus*, and VP5 floor domain from herpesvirus, strengthening the idea of a common evolutionary root. Analogous to the bones of extinct organisms locked in geological layers, we suggest this ancient provirus in the genome of SsP2 provides a “fossil” record of dsDNA viruses (Akita et al., 2007; Schwartz et al., 1996; Tipper, Bencsics, and Coffin, 2005; Yamada and Iwamoto, 2000).

Results

Infection of *Sulfolobus solfataricus* P2 and Particle Discovery

The paucity of information about viruses that infect archaea and placement of archaea at the base of most trees of life has stimulated research in this area. Along these lines, we have been conducting structure/function studies on the crenarchaeal *Sulfolobus* turreted icosahedral virus (STIV)(Fulton et al., 2009; Larson et al., 2006; Rice et al., 2004; Wiedenheft et al., 2005). As part of this work, proteomic analysis of the infection of *S. solfataricus* P2 cells by STIV led to the identification of the differentially expressed hypothetical protein Sso2749 (GenBank ID: [15899464](#)). The 2-D gel spot containing Sso2749 decreased in abundance as determined by Cy-dye based 2-DIGE (12 hr timepoint; 2.7 fold reduction, p-value < 0.01), fig. 2.1A. BLAST sequence analysis revealed that Sso2749 has significant amino acid sequence similarity to a virus-like particle from *Pyrococcus furiosus* (Pf-vlp) (GenBank ID: [146387236](#)) (fig. S1)(Altschul et al., 1997). Previously, Pf-vlp was shown to form virus-like particles of approximately 30 nm as revealed by transmission electron microscopy (TEM) (Namba et al., 2005). Subsequently, the crystal structure revealed that Pf-vlp assembles from 180 copies of protein into a T=3 icosahedral particle with the monomer having significant structural similarity to major capsid protein (MCP) gp5 of HK97 bacteriophage (GenBank ID: [609312](#)) and encapsulin from *Thermatoga maritima* (GenBank ID: [197725260](#))(Sutter et al., 2008). To check for virus-like particles in *S. solfataricus*, fractions of cell lysate were pelleted by ultra-centrifugation onto a sucrose cushion.

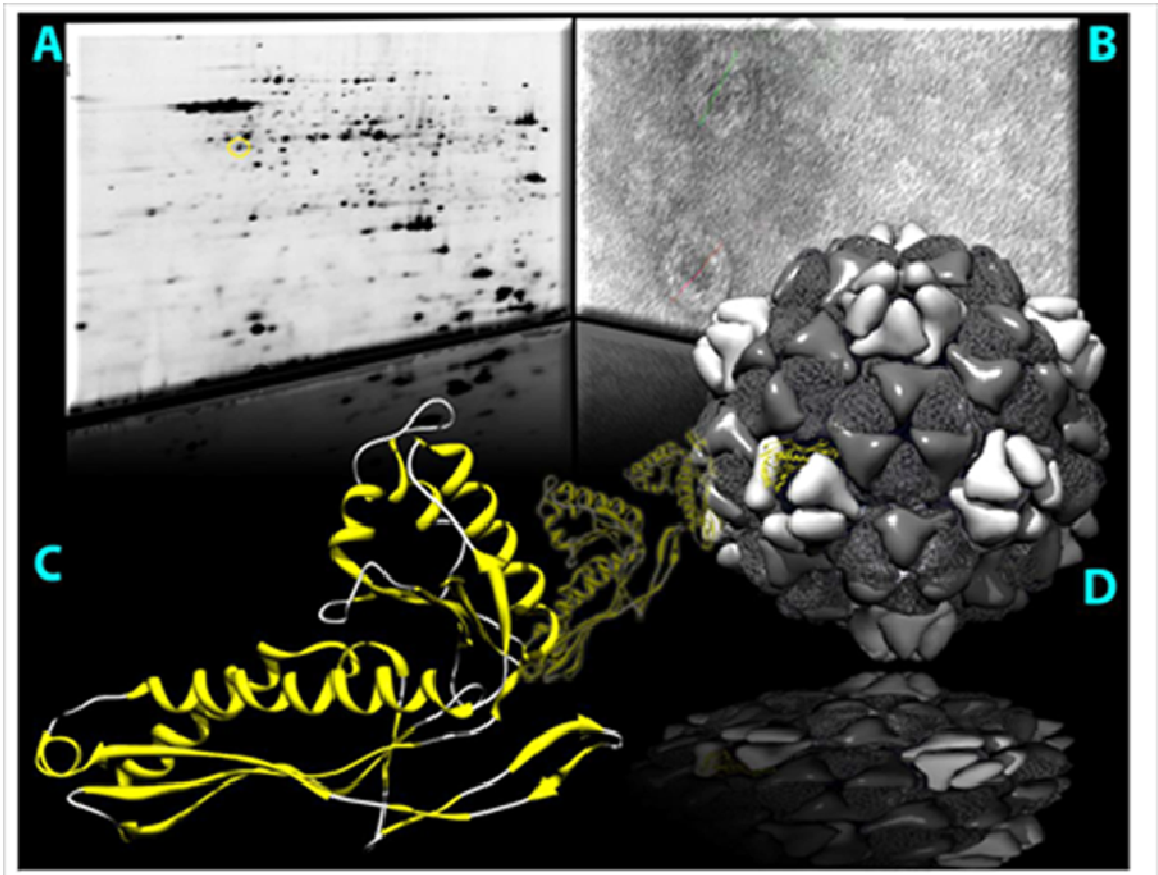


Figure 2.1 Treatment of *S. solfataricus* with STIV Leads to a Down Regulation of Translation of Ss-vlp. (A) 2-DIGE gel image with protein spot for hypothetical protein SSO2749 highlighted. Protein identification was made using in-gel proteolysis followed by LC-MS/MS (B) Electron micrograph of purified rSs-vlp after negative staining with 1.5% uranyl acetate. (C) Monomer for hypothetical protein SSO2749 modeled using 2E0Z PfV (X-ray crystal structure 3.6 Angstrom resolution) (D) Ss-vlp model using 180 copies of modeled SSO2749 with T = 3 icosahedral symmetry based on crystal structure of PfV (Akita, Chong et al. 2007)

The pellet was then analyzed by LCMS and found to contain Sso2749. Subsequent analysis with TEM confirmed the presence of particles (~30-32 nm) similar to *Pf*-vlp (fig. 2.1B). To better investigate the structural and functional properties of Ss-vlp, the gene was cloned into an expression vector for heterologous expression in *E. coli*. Analysis of the soluble protein fraction by SDS-PAGE showed the presence of an intense band at the expected molecular weight of Sso2749 (39kDa)(fig. S2). Analysis of partially purified

Ss-vlp (rSs-vlp) by size exclusion chromatography (SEC) confirmed that the recombinant protein was multimeric, with an elution time consistent with Cowpea chlorotic mottled virus (CCMV), a small RNA virus that has a T=3 capsid. Identity of the protein in the SEC fractions and gel band was confirmed by proteolysis and LC/MS (data not shown). As a final test, negative stain TEM analysis of purified Sso2749 showed particles of the expected size for a T=3 vlp (fig. 2). The majority of the particles appeared to have a hollow center and a thick shell. This is consistent with the images of the Pf-vlp and positioning of the N-terminal region of the gene on the interior of the particle. Purified fractions of rSs-vlp were tested for possible linocin-like bacteriocin activity and for susceptibility to proteolytic cleavage. No bacteriocidal activity was observed when rSs-vlp was added to cultures of *Pyrococcus furiosus*, or *Sulfolobus acidocaldarius*, as judged by optical density growth curves (data not shown). Testing also showed that the particles were resistant to proteolytic cleavage by trypsin and thermolysin.

In an attempt to gain insight into the structure of the Ss-vlp, a bioinformatics approach was used. Sequence-based structure/function prediction using the Phyre and HHpred servers (Hildebrand et al., 2009; Kelley and Sternberg, 2009; Soding, Biegert, and Lupas, 2005) reinforced the homology between Ss-vlp and Pf-vlp showing that they are highly likely to have the same protein fold. Therefore, a 3-D model, using the Pf-vlp crystal structure (PDB ID: [2E0Z](#)) as a template, was constructed for Ss-vlp residues 110-346 with Swiss-Model (Arnold et al., 2006; Guex and Peitsch, 1997; Kiefer et al., 2009) (fig.1C-D). The N-terminal 109 residues of Pf-vlp are not visible in the crystal structure,

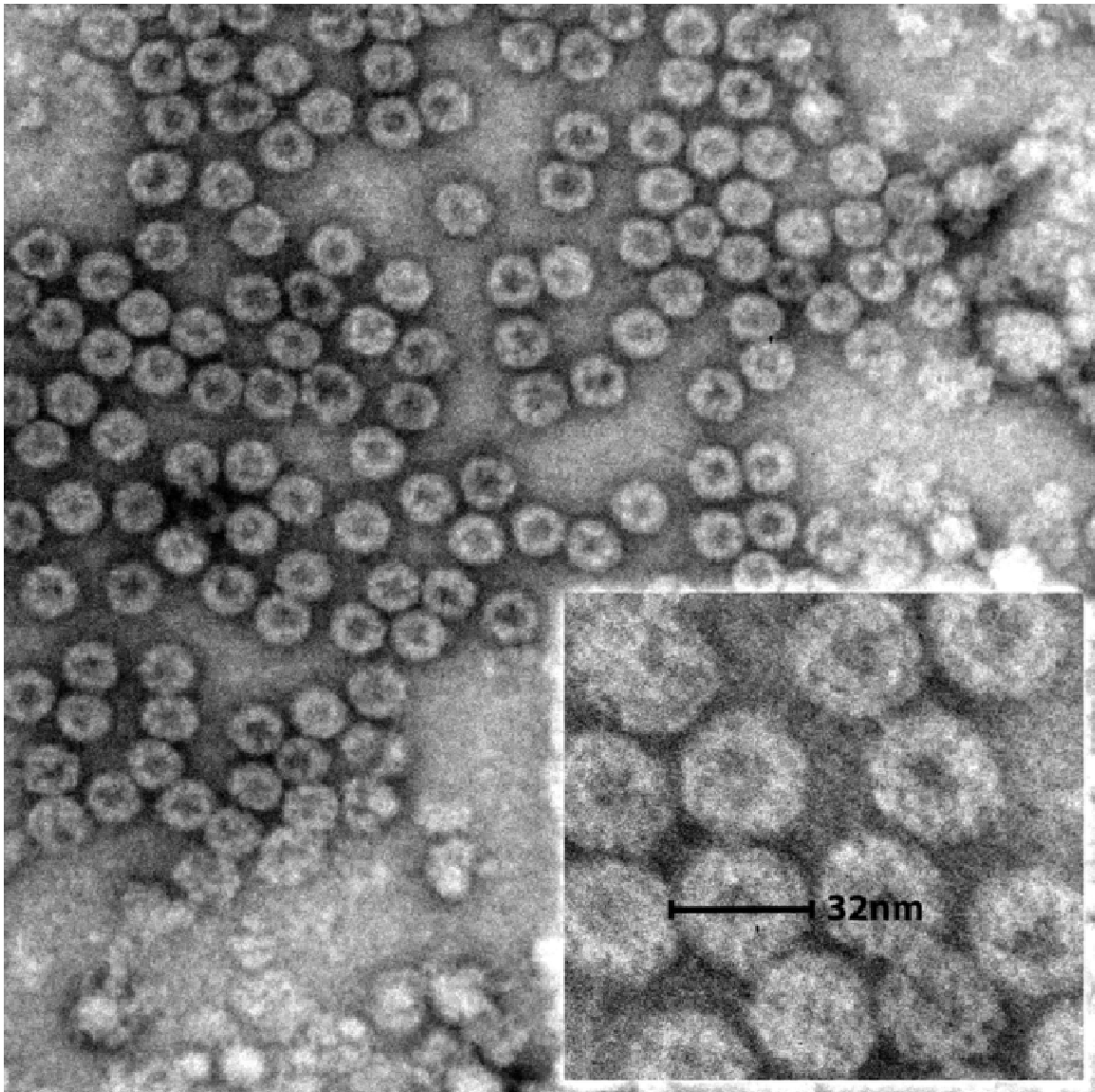


Figure 2.2 Electron Micrographs of Recombinant SsP2-vlps After Negative Staining with 1.5% Uranyl Acetate. Foreground is a close up of the same particles.

but are believed to form a domain that is internal to the protein shell. HHpred analysis of residues 1-109 from Ss-vlp indicates that these residues form a putative ferritin-like DIIRON-carboxylate domain (PDB ID: [1vjx_A](#)). The quality of the Ss-vlp model was then verified using ProSA-web (Wiederstein and Sippl, 2007). The Z-score of the pseudo-atomic model of Ss-vlp (Z-score = -2.96) is within the range calculated for

experimentally determined structures where acceptable models fall between -2 and -11.2 for proteins of ~ 236 aa in length (fig. S3). Although there is no significant amino acid sequence identity between Ss-vlp, encapsulin, and HK97 GP5, structural alignment of these three proteins with Pf-vlp suggest the same protein fold. The overall analysis of Ss-vlp and rSs-vlp shows that the protein assembles into icosahedral capsids similar to tailed dsDNA viruses (Wikoff et al., 2000), *Pf*-vlp (Akita et al., 2007) or encapsulin (Sutter et al., 2008) and strongly suggests that this protein adopts an HK-97 like fold.

Characterization of a Provirus in *S. solfataricus*

The differential expression of Ss-vlp during viral infection combined with structural similarity to a known family of temperate viruses, and previous findings by other groups (Akita et al., 2007; Sutter et al., 2008), led us to hypothesize that these particles were of viral origin. Viruses are known to recombine with host and other viral elements to yield unique genetic mosaics that are capable of promoting horizontal gene transfer (Brussow and Desiere, 2001; Hendrix, 2002; Juhala et al., 2000; Mosig et al., 2001; Rokyta et al., 2006). To determine if Ss-vlp was part of an integrated provirus, the genomic context of the gene was examined. In order to identify other phage-like or provirus elements, iterative BLAST homology-based searches against protein sequences in the non-redundant protein database and dsDNA viruses (taxid: [35237](#)) at the National Center for Biotechnology Information were conducted, using genes in close proximity to Sso2749 as seeds. Profile-profile comparisons are almost always indicative of true functional and structural relationships and are particularly valuable when looking for evolutionary relationships. (Altschul et al., 1997); however protein sequences of viruses

infecting *Sulfolobus* tend to bear little similarity to other known proteins making comparisons more problematic (Fulton et al., 2009; Prangishvili, Stedman, and Zillig, 2001; Prangishvili et al., 2006). Recent comparisons of proteins using discriminative state-of-the-art machine learning algorithms, have found this to be a reliable approach for assessing distant relationships; therefore, the profile based comparisons were complemented by analyses with support vector machines (Dlakic, 2009; Shah, Oehmen, and Webb-Robertson, 2008).

Consistent with a viral origin for SsP2-vlp, proteins with similarity to a phage replicative helicase, and herpesvirus DNA binding protein lie directly 3' of Sso2749 (table 1). The helicase and DNA binding protein genes are in an operon which has been shown to be induced by exposure to ultraviolet light (Frols et al., 2007); UV light is a common trigger for provirus induction (Muche and Levin, 1983; Osterhout et al., 2007; Takebe et al., 1967; Trautner et al., 1980). For homologs found in a proviral region from the genome of *Sulfolobus islandicus* strain Y.G.57.14, all three, virus MCP, helicase, and DNA binding protein are found to as an operon, solidifying their codependence. A tRNA-Thr gene (genome position: 2503123) was then found in a 3' distal region approximately 36 kbp from Ss-vlp (genome position: 2539011). It has been shown that bacterial and archaeal viruses use 5' or 3'-distal regions of tRNA genes as targets for integration. Krupovic *et al*, 2010 reported tRNA-Thr as an integration target for a prophage found in *Halorubrum lacusprofundi*, (ATCC 49239) a haloarchaeal organism. To further support the presence of a provirus, the genomic neighborhood surrounding the

SsP2-protein	SsP2-GI #	Organism / Virus	Protein	GI/Accession	E-value/P-value
cSSO2724	15899440	<i>Bacteriophage Mu</i>	Integrase	PF02914	6e-06
*SSO2726	13816082	<i>Stygiolobus rod-shaped virus</i>	Hypothetical	208605314	6e-07
*SSO2727	13816050	<i>Mycobacterium phage Myrna</i>	gp183	203454746	1e-12
*SSO2728	13816054	<i>Stygiolobus rod-shaped virus</i>	hypothetical	208605314	4e-05
cSSO2730	15899446	<i>Acinetobacter baumannii</i>	hypothetical bacteriophage	169796743	1e-15
cSSO2733	15899448	<i>Desulfuromonas acetoxidans</i>	Phage SP01 DNA polymerase-related protein	95929443	2e-10
cSSO2744	15899459	<i>Bacteriophage</i>	gp37	COG4422	4e-12
cSSO2750	15899465	<i>Phage</i>	DnaB Replicative helicase	TIGR03600	5e-7
*SSO2751	13816082	<i>Caviid herpesvirus</i>	gp44	213159193	9e-03
cSSO2754	15899469	<i>Escherichia coli MS 21-1</i>	tail fiber protein	300940050	3e-05
*SSO2766	15899482	<i>Anguillid herpesvirus</i>	ORF83	28217420	9e-43
cSSO2778	15899493	<i>Lactococcus lactis</i>	phage-nucleotide binding protein	281491928	4e-12
*SSO2782	13816122	<i>Acidianus two-tailed virus</i>	Putative transposase	75750428	7e-165
*SSO2785	13816122	<i>Acidianus two-tailed virus</i>	Putative transposase	75750428	7e-172

Table 2.1 SsP2 protein matches to viral and proviral proteins; a) protein similarity found in dsDNA viral database (taxid: 35237); b) protein similarity found in NCBI non-redundant protein database c) protein classification done with support vector machine corresponding probability scores are given as a p-value and Pfam/TIGR/COG domain accession

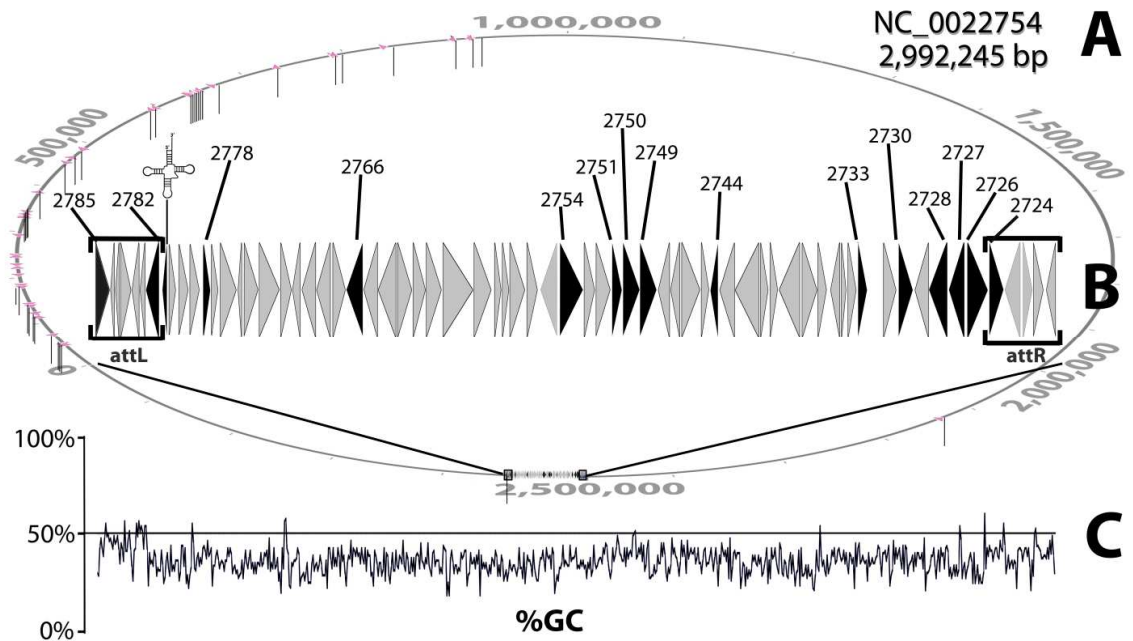


Figure 2.3 A) (Background) Circular Representation of SsP2 Genome with Location of Provirus. vertical lines show location of all structural tRNAs. B) Blowup of provirus genome located between 2,478,239 - 2,539,011, triangles denote open reading frames (ORF), black triangles represent ORF with predicted amino acid similarity to other virus or provirus, boxed regions (attL and attR) are transposon related elements on either side of the provirus. C) %GC graph of the provirus (Drummond AJ, 2010).

structural tRNA and the SsP2 provirus MCP were scanned for other proviral elements (see fig. 3). The identification of transposons flanking both sides of the virus-like region with N-terminal viral integrase allowed us to define likely boundaries of the SsP2-provirus (attL, attR) (fig. 3B). This would credit the integrated provirus with 57 open reading frames in a span of 60.7 kbp (between the attL and attR). The %GC-content in this region is between 27.8% - 41.1% with an overall average of 36%, compared to the SsP2 average of 35% (fig. 3C)(Drummond AJ, 2010). Table 1 shows the proteins with viral or proviral similarity. The SsP2-provirus displays similarity to dsDNA viruses from different domains, exhibiting the mosaicism commonly described in the dsDNA

bacteriophage family (Bamford, Burnett, and Stuart, 2002; Hatfull et al., 2006; Hendrix, 2003; Varani et al., 2008).

Discussion

The high degree of sequence similarity between protein Sso2749 and *Pf*-vlp, combined with the, protein model, and TEM characterization of the particle has allowed us to classify *Ss*-vlp as a homolog of *Pf*-vlp. This establishes a clear connection at the protein level between *Ss*-vlp, encapsulin, and gp5, the MCP of HK97. The genomic context of *Ss*-vlp suggests that this protein was part of a dsDNA-like provirus, placing the HK97 MCP fold in viruses from all three domains of life. This finding is significant, as structural similarity has been put forth as a measure of the viral self (Bamford, Burnett, and Stuart, 2002; Krupovic, Forterre, and Bamford, 2010) and a prognostic indicator of common ancestry (Bamford, Grimes, and Stuart, 2005; Kraft et al., 2004; Maaty et al., 2006). Clearly the observed particles are not of sufficient size to package the identified proviral genome, suggesting that this provirus is no longer functional, and only remains as part of the HK97 virus family fossil record. In the *SsP2* provirus fossil, proteins with similarity to dsDNA bacteriophages can be found side-by-side with proteins that have similarity to herpesvirus proteins. This provides a direct connection between the dsDNA viruses in all three domains and raises the question of whether the sequence similarity found between *SsP2*-pro proteins and their relatives in bacteriophage and herpesvirus is a product of ancestral virus-virus recombination via horizontal gene transfer or a remnant of evolution from a common ancestor.

Pfam classification places *T. maritima* encapsulin, Pf-vlp and Ss-vlp in the Linocin M18 family (PF04454), which currently has members in 69 species. Sutter *et al* 2008, suggests that the icosahedral shell of encapsulins compartmentalize enzymatic activity inside the protein cage to enhance function. It has been proposed by a number of groups that encapsulins may have originated from a viral capsid (Krupovic, Forterre, and Bamford, 2010; Sutter et al., 2008). Because the similarity between encapsulins and viral capsid proteins is only visible at the level of structure and there were no reported cases where encapsulins are associated with viral proteins, it was assumed that the separation occurred a long time ago. The analysis of Sso2749 clearly shows a direct connection of encapsulin proteins with viral origin. The use of the HK97-like protein fold by viruses and host cells points to the versatility of the domain. The domain is also fairly plastic as large extensions can be added without disrupting assembly as in the case of encapsulin. While the connection between encapsulins and a provirus has now been documented, the fundamental question still remains; is the design of host or viral origin? The prevailing view is that in the context of capsids and cages, the HK97 protein is foremost a viral capsid protein. An alternative explanation is that encapsulins were first derived as primitive reaction vessels that enhanced the activity of primitive enzymes. In some cases the enzymes then became linked with the cage forming domain so that packaging and assembly occur in a single step. Then at some later time, the HK97 fold mutated to allow nucleic acid packaging, perhaps by simply assembling around its own mRNA during translation. At this point, the simple system would be ready to go viral.

Materials and Methods

Culturing of *S. solfataricus*

Liquid cultures of *S. solfataricus* (P2), were grown aerobically in DSMZ media 182 (22.78 mM KH_2PO_4 + 18.90 mM $(\text{NH}_4)_2\text{SO}_4$ + 0.81 mM MgSO_4 + 1.7 mM CaCl_2 + 0.2% Yeast Extract) pH adjusted to 2.8 with 6N H_2SO_4 . All cultures were grown in long neck Erlenmeyer flasks at 80 degrees Celsius in shaking oil bath incubators.

Purification of Spherical Particles

Cell cultures of SsP2 were lysed by freezing in liquid Nitrogen and thawing, followed by water bath sonication for 5 minutes. The suspension was then centrifuged at 39,500 x g for 15 minutes to remove insoluble cellular debris from the supernatant. The supernatant was then added to a 20% sucrose cushion suspended in 50 mM Phosphate buffer with 9 g/L NaCl and centrifuged for 4 hours at 150,000 x g in an ultracentrifuge. The resultant pellet formed was resuspended in 150 ul, 50 mM Phosphate buffer with 9 g/L NaCl and subjected to SDS-PAGE (4-20% Precise protein gels, Pierce), for 1 hr at 65 mA. Gel was then washed and stained using Coomassie Brilliant Blue (CBB). Protein spots of interest were excised from the gels, washed, in-gel reduced and S-alkylated, followed by digestion with porcine trypsin (Promega) overnight at 37°C. The Peptides eluted from the in-gel digests were transferred to a new tube prior to LC/MS/MS. Mass analysis was on an integrated Agilent 1100 liquid chromatography–mass-selective detection (LC-MSD) trap (XCT-Ultra 6330) controlled with ChemStation LC 3D (Rev A.10.02). The Agilent XCT-Ultra ion trap mass spectrometer is fitted with an Agilent

1100 CapLC and nano-LC sprayer under the control of MSD trap control version 5.2 Build no. 63.8 (Bruker Daltonic GmbH). Injected samples were first trapped and desalted on the Zorbax 300SB-C18 Agilent HPLC-Chip enrichment column (4 mm with 40 nl volume packed with 5 μ m Zorbax 300SB-C18 particles; Agilent) for 3 min with 0.1% formic acid delivered by the auxiliary pump at 4 μ l/min. The peptides were then reverse eluted from a enrichment column and loaded onto an analytical capillary column (43 mm x 75 μ m inside diameter, packed with 5 μ m Zorbax 300SB-C18 particles) connected in-line to the mass spectrometer with a flow of 600 nl/min. Peptides were eluted with a 5 to 90% acetonitrile gradient over 16 min. Data-dependent acquisition of collision induced dissociation tandem mass spectrometry (MS/MS) was utilized. Parent ion scans were run over the m/z range of 400 to 2,200 at 24,300 m/z -s. MGF compound list files were used to query an in-house database using Biotoools software version 2.2 (Bruker Daltonic) with 0.5 Da MS/MS ion mass tolerance. Protein identification was accomplished by database search using MASCOT (Matrix science, London, UK). Protein identification was considered positive when the minimal protein score of greater than 100 indicated identity or extensive homology.

Cloning

The PCR primers were added to attB sites to facilitate ligase-free cloning using the Gateway II system (Invitrogen) and a Shine-Delgarno sequence to facilitate efficient translation. The internal forward and reverse primers were 5'-ATG CAT CAC CAT CAC CAT CAC ATG AGT GAG GAA AGA ATT CCA TTA ATA GG-3' and 5' - GTA CAA GAA AGC TGG GTC CTA TTA TGT TGT TTT AAC TTC TTT ATA AGT GAA

CCA C-3', respectively. Oligonucleotides used for the external forward and reverse primers were 5'- GGG GAC AAG TTT GTA ACA AAA AAG CAG GCT TCG AAG GAG ATA GAA CC-3' and 5'-GGG GAC CAC TTT GTA CAA GAA AGC TGG GTC CTA-3' respectively. The resulting entry clone was determined using ABI BigDye terminator cycle sequencing and is mutation free. This construct was then inserted into pDEST14 (Invitrogen) destination vector, yielding an pEXP14-Sso2749 expression vector, for protein expression in *Escherichia coli* (*E. coli*).

Expression and Purification

pEXP14-Sso2749 was transformed into *E. coli* BL21- CodonPlus(DE3)-RIL (Stratagene), and two colonies were used to inoculated two 2 ml cultures of Luria-Bertani (LB) media and grown overnight at 37 C. 2 ml cultures were used inoculate 100 ml of LB media. All media contained 100 ug/ml ampicillin and 35 ug/ml chloramphenicol. Cells were grown to an optical density of 0.6 at 600 nm then protein expression was induced via addition of 1 mM isopropyl-D-thiogalactopyranoside (IPTG). After 6 hours of growth, cells were collected by centrifugation (IEC PR-7000 M) at 6,000 x g for 15 minutes and pellets were stored at -80 C (8).

For purification, cell pellets were thawed and resuspended in lysis buffer (20mM Tris-HCl, pH 8.0, 400 mM NaCl and 1mM EDTA) at 10 ml/g of wet cell pellet mass. Phenylmethylsulphonyl-fluoride (PMSF, 0.1 mM) was added to the cell suspension and cells were lysed by passage through a microfluidizer (Microfluidics Corporation, Newton, MA). Cell lysate was centrifuged and soluble fraction was discarded and the insoluble fraction was washed with lysis buffer and pelleted by centrifugation. The

insoluble fraction was solubilized in 10 ml solubilization buffer (50 mM Tris-HCl pH 8.0, 24 mM NaCl, 1 mM KCl, 8 M urea, and 5 mM β -mercaptoethanol) and the remaining insoluble material was removed by centrifugation. The supernatant was dialyzed in two steps: first, dialyzed against refolding buffer (50 mM Tris-HCl pH 8.0, 24 mM NaCl, 1 mM KCl) containing 4M urea then dialyzed against refolding buffer only. Initial purification was done by size-exclusion chromatography on Sephacryl S-500 (Amersham Pharmacia). Further protein was purified to homogeneity by ion exchange (mono-Q) and size exclusion (Sephacryl S-500) chromatographic techniques.

Electron Microscopy

For negative stain TEM samples of purified SsV was spotted onto copper grid. The protein was set to absorb for 2 minutes before removing excess water by absorption on a piece of filter paper. Uranyl acetate (1.5%) was added to grid for 30 seconds at which point excess fluid was wicked away with piece of Whatman filter paper for a second time, and then allowed to dry. Stained SsV on copper grid was viewed with a LEO 912AB TEM and photographed with a Proscan 2,048- by 2,048-pixel charge-coupled- device camera.

Proteolysis

Purified Sso2749 was exposed to trypsin for 1 hr at 37 C and thermolysin for 1hr at 70 C. After which each concentration was removed and subjected to SDS-PAGE (4-20% Precise protein gels, Pierce) for 1 hr at 65 mA. Gel was then washed and stained using Coomassie Brilliant Blue (CBB).

Identification of Putative Proviruses

Putative provirus elements related to tailed bacterial and archaeal viruses were identified by homology-based searches against the non-redundant protein database at NCBI. Analysis of putative viral protein sequences Protein sequences of the archaeal viruses and putative proviruses were analyzed using BLASTP or PSI-BLAST search against (i) a complete nonredundant protein database at NCBI and (ii) a viral protein database at NCBI. Additionally protein sequences were investigated using Hidden Markov Models and support vector machines as described by Dlakic et al 2009. Conserved protein domains were identified using searches against the Conserved Domain Database at NCBI and the Pfam database at Sanger.

Prediction, Modeling, and Verification of Pseudoatomic Models

MCP (Ss-vlp) sequence of putative tailed archaeal provirus was submitted to the Phyre structure prediction server for fold prediction and identification of structural homologues. All X-ray structures that were identified as structural homologues with significant scores (scores above 50 are considered significant) were used as templates for the structural modeling of the SsP2 SsV. Sequence alignment of the target MCP was retrieved from SWISS Model-workspace and used to build three-dimensional model with version 4 of the Deepview program. The quality of the model was then assessed with ProSA-web. Pairwise sequence alignment used to generate the model presented in Fig. 5 can be found in Supplementary Information.

References

- Akita, F., Chong, K. T., Tanaka, H., Yamashita, E., Miyazaki, N., Nakaishi, Y., Suzuki, M., Namba, K., Ono, Y., Tsukihara, T., and Nakagawa, A. (2007). The crystal structure of a virus-like particle from the hyperthermophilic archaeon *Pyrococcus furiosus* provides insight into the evolution of viruses. *Journal of Molecular Biology* **368**(5), 1469-1483.
- Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**(17), 3389-3402.
- Arnold, K., Bordoli, L., Kopp, J., and Schwede, T. (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**(2), 195-201.
- Bamford, D. H. (2003). Do viruses form lineages across different domains of life? *Research in Microbiology* **154**(4), 231-236.
- Bamford, D. H., Burnett, R. M., and Stuart, D. I. (2002). Evolution of viral structure. *Theoretical Population Biology* **61**(4), 461-470.
- Bamford, D. H., Grimes, J. M., and Stuart, D. I. (2005). What does structure tell us about virus evolution? *Current Opinion in Structural Biology* **15**(6), 655-663.
- Brussow, H., Canchaya, C., and Hardt, W. D. (2004). Phages and the evolution of bacterial pathogens: From genomic rearrangements to lysogenic conversion. *Microbiology and Molecular Biology Reviews* **68**(3), 560.
- Brussow, H., and Desiere, F. (2001). Comparative phage genomics and the evolution of Siphoviridae: insights from dairy phages. *Molecular Microbiology* **39**(2), 213-222.
- Campbell, A. (2003). Prophage insertion sites. *Research in Microbiology* **154**(4), 277-282.
- Chibani-Chennoufi, S., Bruttin, A., Dillmann, M. L., and Brussow, H. (2004). Phage-host interaction: an ecological perspective. *Journal of Bacteriology* **186**(12), 3677-3686.
- Dlakic, M. (2009). HHsvm: fast and accurate classification of profile-profile matches identified by HHsearch. *Bioinformatics* **25**(23), 3071-3076.

- Drummond AJ, A. B., Buxton S, Cheung M, Cooper A, Heled J, Kearse M, Moir R, Stones-Havas S, Sturrock S, Thierer T, Wilson A (2010). Geneious v5.1.
- Dunigan, D. D., Fitzgerald, L. A., and Van Etten, J. L. (2006). Phycodnaviruses: A peek at genetic diversity. *Virus Research* **117**(1), 119-132.
- Forde, S. E., Thompson, J. N., Holt, R. D., and Bohannan, B. J. M. (2008). Coevolution drives temporal changes in fitness and diversity across environments in a bacteria-bacteriophage interaction. *Evolution* **62**(8), 1830-1839.
- Frols, S., Gordon, P. M. K., Panlilio, M. A., Duggin, I. G., Bell, S. D., Sensen, C. W., and Schleper, C. (2007). Response of the hyperthermophilic archaeon *Sulfolobus solfataricus* to UV damage. *Journal of Bacteriology* **189**(23), 8708-8718.
- Fulton, J., Bothner, B., Lawrence, M., Johnson, J. E., Douglas, T., and Young, M. (2009). Genetics, biochemistry and structure of the archaeal virus STIV. *Biochemical Society Transactions* **37**, 114-117.
- Guex, N., and Peitsch, M. C. (1997). SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**(15), 2714-2723.
- Hatfull, G. F. (2008). Bacteriophage genomics. *Current Opinion in Microbiology* **11**(5), 447-453.
- Hatfull, G. F., Pedulla, M. L., Jacobs-Sera, D., Cichon, P. M., Foley, A., Ford, M. E., Gonda, R. M., Houtz, J. M., Hryckowian, A. J., Kelchner, V. A., Namburi, S., Pajcini, K. V., Popovich, M. G., Schleicher, D. T., Simanek, B. Z., Smith, A. L., Zdanowicz, G. M., Kumar, V., Peebles, C. L., Jacobs, W. R., Lawrence, J. G., and Hendrix, R. W. (2006). Exploring the mycobacteriophage metaproteome: Phage genomics as an educational platform. *Plos Genetics* **2**(6), 835-847.
- Hendrix, R. W. (2002). Bacteriophages: Evolution of the majority. *Theoretical Population Biology* **61**(4), 471-480.
- Hendrix, R. W. (2003). Bacteriophage genomics. *Current Opinion in Microbiology* **6**(5), 506-511.
- Heymann, J. B., Cheng, N. Q., Newcomb, W. W., Trus, B. L., Brown, J. C., and Steven, A. C. (2003). Dynamics of herpes simplex virus capsid maturation visualized by time-lapse cryo-electron microscopy. *Nature Structural Biology* **10**(5), 334-341.
- Hildebrand, A., Remmert, M., Biegert, A., and Soding, J. (2009). Fast and accurate automatic structure prediction with HHpred. *Proteins-Structure Function and Bioinformatics* **77**, 128-132.

- Jiang, S. C., and Paul, J. H. (1998). Gene transfer by transduction in the marine environment. *Applied and Environmental Microbiology* **64**(8), 2780-2787.
- Juhala, R. J., Ford, M. E., Duda, R. L., Youlton, A., Hatfull, G. F., and Hendrix, R. W. (2000). Genomic sequences of bacteriophages HK97 and HK022: Pervasive genetic mosaicism in the lambdoid bacteriophages. *Journal of Molecular Biology* **299**(1), 27-51.
- Kelley, L. A., and Sternberg, M. J. E. (2009). Protein structure prediction on the Web: a case study using the Phyre server. *Nature Protocols* **4**(3), 363-371.
- Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L., and Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research* **37**, D387-D392.
- Kraft, P., Kummel, D., Oeckinghaus, A., Gauss, G. H., Wiedenheft, B., Young, M., and Lawrence, C. M. (2004). Structure of D-63 from *Sulfolobus* spindle-shaped virus 1: Surface properties of the dimeric four-helix bundle suggest an adaptor protein function. *Journal of Virology* **78**(14), 7438-7442.
- Krupovic, M., Forterre, P., and Bamford, D. H. (2010). Comparative Analysis of the Mosaic Genomes of Tailed Archaeal Viruses and Proviruses Suggests Common Themes for Virion Architecture and Assembly with Tailed Viruses of Bacteria. *Journal of Molecular Biology* **397**(1), 144-160.
- Larson, E. T., Reiter, D., Young, M., and Lawrence, C. M. (2006). Structure of A197 from *Sulfolobus* turreted icosahedral virus: a crenarchaeal viral glycosyltransferase exhibiting the GTA fold. *Journal of Virology* **80**(15), 7636-7644.
- Lin, G. G., and Li, J. M. (2009). Sequence Identity between the Genomes of Humans and Viruses. *Intervirology* **52**(4), 196-200.
- Maaty, W. S. A., Ortmann, A. C., Dlakic, M., Schulstad, K., Hilmer, J. K., Liepold, L., Weidenheft, B., Khayat, R., Douglas, T., Young, M. J., and Bothner, B. (2006). Characterization of the archaeal thermophile *Sulfolobus* turreted icosahedral virus validates an evolutionary link among double-stranded DNA viruses from all domains of life. *Journal of Virology* **80**(15), 7625-7635.
- Moreira, D. (2000). Multiple independent horizontal transfers of informational genes from bacteria to plasmids and phages: implications for the origin of bacterial replication machinery. *Molecular Microbiology* **35**(1), 1-5.
- Mosig, G., Gewin, J., Luder, A., Colowick, N., and Vo, D. (2001). Two recombination-dependent DNA replication pathways of bacteriophage T4, and their roles in mutagenesis and horizontal gene transfer. *Proceedings of the National Academy of Sciences of the United States of America* **98**(15), 8306-8311.

Muche, A. A., and Levin, R. E. (1983). Repression of Uv Induction of Lambda-Prophage by Caffeine. *Mutation Research* **119**(1), 1-6.

Namba, K., Hagiwara, K., Tanaka, H., Nakaishi, Y., Chong, K. T., Yamashita, E., Armah, G. E., Ono, Y., Ishino, Y., Omura, T., Tsukihara, T., and Nakagawa, A. (2005). Expression and molecular characterization of spherical particles derived from the genome of the hyperthermophilic euryarchaeote *Pyrococcus furiosus*. *Journal of Biochemistry* **138**(2), 193-199.

Osterhout, R. E., Figueroa, I. A., Keasling, J. D., and Arkin, A. P. (2007). Global analysis of host response to induction of a latent bacteriophage. *Bmc Microbiology* **7**, -.

Prangishvili, D., Stedman, K., and Zillig, W. (2001). Viruses of the extremely thermophilic archaeon *Sulfolobus*. *Trends in Microbiology* **9**(1), 39-43.

Prangishvili, D., Vestergaard, G., Haring, M., Aramayo, R., Basta, T., Rachel, R., and Garrett, R. A. (2006). Structural and genomic properties of the hyperthermophilic archaeal virus ATV with an extracellular stage of the reproductive cycle. *Journal of Molecular Biology* **359**(5), 1203-1216.

Prangishvili, D., Garrett, R. A., and Koonin, E. V. (2006). Evolutionary genomics of archaeal viruses: Unique viral genomes in the third domain of life. *Virus Research* **117**(1), 52-67. Rao, V. B., and Feiss, M. (2008). The Bacteriophage DNA Packaging Motor. *Annual Review of Genetics* **42**, 647-681.

Rice, G., Tang, L., Stedman, K., Roberto, F., Spuhler, J., Gillitzer, E., Johnson, J. E., Douglas, T., and Young, M. (2004). The structure of a thermophilic archaeal virus shows a double-stranded DNA viral capsid type that spans all domains of life. *Proceedings of the National Academy of Sciences of the United States of America* **101**(20), 7716-7720.

Rokyta, D. R., Burch, C. L., Caudle, S. B., and Wichman, H. A. (2006). Horizontal gene transfer and the evolution of microvirid coliphage genomes. *Journal of Bacteriology* **188**(3), 1134-1142.

Schwartz, D. H., Viscidi, R., Laeyendecker, O., Song, H. F., Ray, S. C., and Michael, N. (1996). Predominance of defective proviral sequences in an HIV plus long-term non-progressor. *Immunology Letters* **51**(1-2), 3-6.

Shah, A. R., Oehmen, C. S., and Webb-Robertson, B. J. (2008). SVM-HUSTLE - an iterative semi-supervised machine learning approach for pairwise protein remote homology detection. *Bioinformatics* **24**(6), 783-790.

Soding, J., Biegert, A., and Lupas, A. N. (2005). The HHpred interactive server for protein homology detection and structure prediction. *Nucleic Acids Research* **33**, W244-W248.

Sutter, M., Boehringer, D., Gutmann, S., Guenther, S., Prangishvili, D., Loessner, M. J., Stetter, K. O., Weber-Ban, E., and Ban, N. (2008). Structural basis of enzyme encapsulation into a bacterial nanocompartment. *Nature Structural & Molecular Biology* **15**(9), 939-947.

Suttle, C. A. (2005). Viruses in the sea. *Nature* **437**(7057), 356-361.

Takebe, H., Ichikawa, H., Iwo, K., and Kondo, S. (1967). Phage Induction by Ultraviolet Radiation in Strains of Escherichia Coli Possessing and Lacking Dark Repair Capacity. *Virology* **33**(4), 638-&.

Tipper, C. H., Bencsics, C. E., and Coffin, J. A. (2005). Characterization of Hortulanus endogenous murine leukemia virus, an endogenous provirus that encodes an infectious murine leukemia virus of a novel subgroup. *Journal of Virology* **79**(13), 8316-8329.

Trautner, T. A., Pawlek, B., Gunthert, U., Canosi, U., Jentsch, S., and Freund, M. (1980). Restriction and Modification in Bacillus-Subtilis - Identification of a Gene in the Temperate Phage Spbeta Coding for a Bsur Specific Modification Methyltransferase. *Molecular & General Genetics* **180**(2), 361-367.

Varani, A. D., Souza, R. C., Nakaya, H. I., de Lima, W. C., de Almeida, L. G. P., Kitajima, E. W., Chen, J., Civerolo, E., Vasconcelos, A. T. R., and Van Sluys, M. A. (2008). Origins of the Xylella fastidiosa Prophage-Like Regions and Their Impact in Genome Differentiation. *Plos One* **3**(12), -.

Wiedenheft, B., Mosolf, J., Willits, D., Yeager, M., Dryden, K. A., Young, M., and Douglas, T. (2005). An archaeal antioxidant: Characterization of a Dps-like protein from Sulfolobus solfataricus. *Proceedings of the National Academy of Sciences of the United States of America* **102**(30), 10551-10556.

Wiederstein, M., and Sippl, M. J. (2007). ProSA-web: interactive web service for the recognition of errors in three-dimensional structures of proteins. *Nucleic Acids Research* **35**, W407-W410.

Wikoff, W. R., Liljas, L., Duda, R. L., Tsuruta, H., Hendrix, R. W., and Johnson, J. E. (2000). Topologically linked protein rings in the bacteriophage HK97 capsid. *Science* **289**(5487), 2129-2133.

Woese, C. R. (2000). Interpreting the universal phylogenetic tree. *Proceedings of the National Academy of Sciences of the United States of America* **97**(15), 8392-8396.

Yamada, T., and Iwamoto, A. (2000). Comparison of proviral accessory genes between long-term nonprogressors and progressors of human immunodeficiency virus type 1 infection. *Archives of Virology* **145**(5), 1021-1027.

CHAPTER THREE

EXPANDING THE PARADIGM OF THIOL REDOX
IN THE THERMOPHILIC ROOT OF LIFEContribution of Authors and Co-Authors

Manuscript in Chapter 3

Author: Joshua Heinemann

Contributions: Culturing of *S. solfataricus*; labeled proteins; extracted and ran untargeted metabolomics; small molecule labeling; quantified NADPH; did bioinformatic protein comparisons; Wrote manuscript.

Co-Author: Timothy Hamerly

Contributions: Culturing of *S. solfataricus*; ran standard on MS; designed part of figure; edited manuscript.

Co-Author: Walid S. Maaty

Contributions: Provided valuable insight; edited manuscript

Co-Author: Navid Movahed

Contribution: Provided valuable insight; edited manuscript

Co-Author: Joseph D. Steffens

Contribution: Developed protein labeling method

Co-Author: Benjamin D. Reeves

Contribution: Synthesized z-dye; Provided valuable insight; edited manuscript.

Co-Author: Jonathan K. Hilmer

Contributions: Provided valuable insight; edited manuscript.

Contribution of Authors and Co-Authors Continued

Co-Author: Jesse Therien

Contributions: Provided valuable insight; edited manuscript.

Co-Author: Paul A. Grieco

Contributions: procured funding.

Co-Author: John W. Peters

Contributions: Provided valuable insight; edited manuscript; procured funding.

Co-Author: Brian Bothner

Contributions: Provided valuable insight; wrote manuscript, procured funding.

Manuscript Information Page

Authors: Joshua Heinemann, Timothy Hamerly, Walid S. Maaty, Navid Movahed, Joseph D. Steffens, Benjamin D. Reeves, Jonathan K. Hilmer, Jesse Therien, Paul A. Grieco, John W. Peters and Brian Bothner

Journal: Biochimica et Biophysica Acta

Status of the manuscript:

☐ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☐ Accepted by a peer-reviewed journal

☒ Published in a peer-reviewed journal

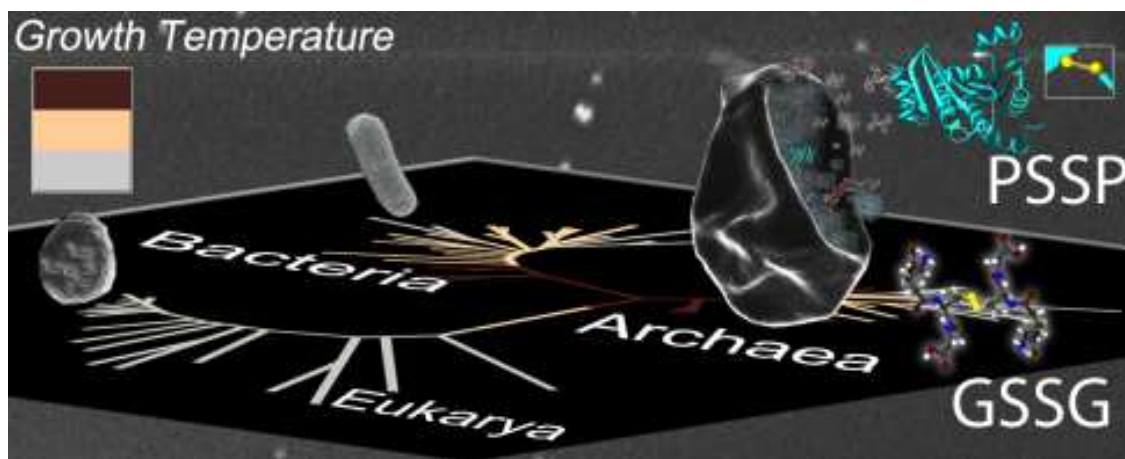
Publisher: Elsevier

Issue in which manuscript appears: Volume 1840, 80-85 (2014)

EXPANDING THE PARADIGM OF THIOL REDOX IN
THE THERMOPHILIC ROOT OF LIFE

Joshua Heinemann^{*}, Timothy Hamerly^{*}, Walid S. Maaty^{*}, Navid Movahed^{*}, Joseph D. Steffens^{*}, Benjamin D. Reeves^{*}, Jonathan K. Hilmer[†], Jesse Therien^{*‡}, Paul A. Grieco^{*}, John W. Peters^{*‡} and Brian Bothner^{**†}

Department of Chemistry and Biochemistry^{*}, Mass Spectrometry, Proteomics and Metabolomics Core Facility[†], Thermal Biology Institute[‡], Montana State University, Bozeman, MT 59717



Abstract

Background: The current paradigm of intracellular redox chemistry maintains that cells establish a reducing environment maintained by a pool of small molecule and protein thiol to protect against oxidative damage. This strategy is conserved in mesophilic organisms from all domains of life, but has been confounded in thermophilic organisms where evidence suggests intracellular proteins have abundant disulfides.

Methods: Chemical labeling and 2-dimensional gel electrophoresis were used to capture disulfide bonding in the proteome of the model thermophile *Sulfolobus solfataricus*. The redox poise of the metabolome was characterized using both chemical labeling and untargeted liquid chromatography mass spectrometry. Gene annotation was undertaken using support vector machine based pattern recognition.

Results: Proteomic analysis indicated the intracellular protein thiol of *S. solfataricus* was primarily in the disulfide form. Metabolic characterization revealed a lack of reduced small molecule thiol. Glutathione was found primarily in the oxidized state (GSSG), at relatively low concentration. Combined with genetic analysis, this evidence shows that pathways for synthesis of glutathione do exist in the archaeal domain.

Conclusions: In observed thermophilic organisms, thiol abundance and redox poise suggest this system is not directly utilized for protection against oxidative damage. Instead, a more oxidized intracellular environment promotes disulfide bonding, a critical adaptation for protein thermostability.

General Significance: Based on the placement of thermophilic archaea close to the last universal common ancestor in rRNA phylogenies, we hypothesize thiol-based redox

systems are derived from metabolic pathways originally tasked with promoting protein stability.

Introduction

Intracellular oxidation-reduction potential must be carefully balanced to support oxidative metabolic processes, while protecting from detrimental effects of free radical damage. It is generally accepted that the cytoplasm is kept in a reducing (electron rich) state, with small variations between different cell types (Sevier et al. 2002). The cytoplasmic redox potential is typically quantified by measuring relative amounts of protein disulfide (PSSP) to thiol (PSH) in equilibrium with compounds of known redox potential, such as oxidized (GSSG) and reduced glutathione (GSH) (Hansen et al. 2009, Cumming et al. 2004). In the cytoplasm of eukaryotic and most gram negative prokaryotes, GSH and PSH are the predominant forms and exist at significant concentrations (Smirnova et al. 2005, Pastore et al. 2003, Fahey et al. 2001). Whereas, GSSG and PSSP are primarily present in more oxidizing environments, such as extracellular space, mitochondria (Chakravarthi et al. 2006, Fratelli et al. 2005, Pedone et al. 2006, Derman et al. 1991) or as catalytic intermediates.

In contrast, structural and genomic evidence suggests thermophiles and their viruses favor higher ratios of PSSP to PSH. The abundance of PSSP is proposed to be an adaptation that increases protein stability at elevated temperatures (Mallick et al. 2002, Beeby et al. 2005). The genomic analyses show this adaptation is exclusive to thermophilic organisms, with the highest proposed PSSP/PSH ratios observed in the

hyperthermophilic Crenarchaea (Jorda et al. 2011, She et al. 2001, Menon et al. 2008, Larson et al. 2007, Lawrence et al. 2009). *Sulfolobus solfataricus* is a hyperthermophilic Crenarchaeon living optimally at 80°C and a pH of 3, and has been adopted as a model for study of high temperature adaptation (Nunn et al. 2010, Lambie et al. 2005, Albers et al. 2009, Heinemann et al. 2011, Maaty et al. 2009, Maaty et al. 2012). Previous studies of *S. solfataricus* and phylogenetically related thermophiles failed to find glutathione (Hummel et al. 2005, Allocati et al. 2012), raising the question, how do these organisms regulate intracellular redox and protein disulfide formation? Organisms without glutathione typically have an alternative small molecule thiol fulfilling this role (Allocati et al. 2012, Malki et al. 2009); however both proteomic and genomic evidence have failed to provide evidence of common thiol-biosynthesis pathways in *S. solfataricus* and its thermophilic relatives (She et al. 2001, Hummel et al. 2005).

The maintenance of thiol redox homeostasis is fundamentally important to an organism's ability to protect against oxidative damage. PSSP/PSH and GSSG/GSH ratios are often used as a proxy for estimating intracellular redox state and tracking changes. *S. Solfataricus* utilizes a novel enzyme system for regulating oxidative stress (Maaty et al. 2009, Weidenheft et al. 2005), but no evidence supports the role of intracellular thiol in this process. Herein we provide evidence that both the metabolites and proteins of *S. solfataricus* are tuned to operate in a noncanonical redox environment. We find *S. solfataricus* does utilize glutathione, although primarily in the oxidized form. Because the balance of oxidized to reduced protein and small molecule thiol is tilted toward oxidized PSSP and GSSG, this strongly suggests a more oxidized intracellular

environment. Together these results characterize a phenotype which lies outside the traditional paradigm, and provides an expanded model of redox homeostasis in thermophilic organisms.

Materials and Methods

Growth of Microorganisms

S. solfataricus P2 (ATCC) was grown aerobically in liquid DSMZ media 182 (22.78 mM KH_2PO_4 , 18.90 mM $(\text{NH}_4)_2\text{SO}_4$, 0.81 mM MgSO_4 , 1.7 mM CaCl_2 , 0.2% Yeast Extract)(Media 1), or with the addition of 0.1% glucose as carbon source (Media 2), pH adjusted to 2.8 with 6M H_2SO_4 . Media was analyzed to ensure no glutathione or small molecule contaminant affected results. Batch cultures of *S. solfataricus* were grown with media in long neck Erlenmeyer flasks at 80°C. One liter of media was inoculated with 10 ml of log phase (OD_{650} 0.33) culture and divided evenly between ten, 1-liter long neck flasks. At (OD_{650} , 0.35), 50 mls of each 0.25 liter culture was removed and placed in a 50 ml falcon tube as growth controls. Cells were collected at exponential, late exponential and stationary growth phases for metabolite extraction. Liquid cultures of recombinant *Escherichia coli* (Strain: BL21-D3) were grown in Luria-Bertani media (1% Peptone, 0.5% Yeast Extract and 1% NaCl, pH = 7) in 1L shaker flasks at 37°C. For oxidative stress experiment, H_2O_2 was administered to a final concentration of 30 μM (n = 10). One liter of media was inoculated with 10 ml of log phase (OD_{650} 0.33) *S. solfataricus* culture and divided evenly between four, 1-liter long neck flasks. At (OD_{650} , 0.35), 50mls of each 0.25 liter culture was removed and placed in a 50 ml falcon tube as

stressed growth controls. An additional 50 ml aliquot was collected from each culture at 30 min post H₂O₂ inoculation and used for metabolite extraction.

Zdye Maleimide Probe for Thiol Labeling

Cells were lysed with combination of freeze/thaw cycles. Protein sample extracts were prepared from cell suspensions in ice cold, phosphate buffer solution (pH = 6.5, 1mM EDTA). All buffers were degassed and kept on ice during extraction to help ensure lack of thiol/disulfide exchange. Proteins were purified and concentrated with 5-fold volume of cold acetone (-80 °C). Protein concentration was measured with RC/DC Protein Assay Kit (Bio-Rad). For each experiment, 50 ug of protein was used. Three replicates were reduced using 20mM tributylphosphine (Sigma) for 20 min at room temperature, alongside three unreduced controls. Dye labeling was performed using a fluorescent Zdye (ZB-M LC-01-56 , and BDR-III-172) coupled to a maleimide, in a method modified from (Maaty et al. 2012). Dye was added to a final concentration of 5 µM in PBS, pH 6.5, for 20 min at room temperature. Reactions were quenched with 5-fold excess acetone (-80°C). Samples were centrifuged at 15,000xg and the protein pellet was resuspended in 2D gel-loading buffer containing 40 mM DTT. Total protein internal standards were labeled with CyDyes using minimal labeling methods according to the manufacturer's protocol (GE Healthcare). 50 µg of protein extract was labeled separately at 0 °C in the dark for 30 min with 400 pmoles of N-hydroxysuccinimide esters of cyanine dyes (Cy5) dissolved in 99.8% DMF (Sigma). Labeling reactions were quenched by the addition of 1µl of a 10 mM L-lysine solution (Sigma) and left on ice for 10 min. Cy5, reduced and unreduced Zdye samples(ZB-M LC-01-56) were combined

appropriately and mixed with rehydration buffer (7M urea, 2 M thiourea, 4% CHAPS). 2-DE was performed as described elsewhere (23) using precast IPG strips (pH 3-11 NL, 24 cm length; GE Healthcare) in the first dimension (IEF). Labeled samples were combined with a maximum of 450 μ l of rehydration buffer (7M urea, 2M thiourea, 4% CHAPS, 0.5% IPG buffer pH = 3-11 NL, 40mM DTT, and a trace of bromophenol blue) and loaded onto IPG strips. 150 μ g of protein was loaded onto each strip and IEF was carried out with the IPGPhor II (GE Healthcare). Focusing was carried out at 20 °C with a maximum of 50 μ A/strip. Active rehydration was achieved by applying 50V for 12 h. This was followed by a stepwise progression of 500V for 500 Vhr, gradient ramp from 500 to 1000V for 1 h, gradient ramp from 3000-5000V for 1 h, gradient ramp from 5000-8000V for 1 h, the 8000V constant for 44000 Vhr. After IEF separation, the strips were equilibrated twice for 15 min with 50 mM Tris-HCl, 6 M Urea, 30% glycerol, 2% SDS and a trace of bromophenol blue, pH 8.8). The first equilibration solution contained 65 mM DTT, and the second 150 mM iodoacetamide. The strips were sealed on the top of the gels using a sealing solution (0.75% agarose in SDS-Tris HCl buffer). The second-dimension SDS-PAGE was performed in a Dalt II (GE Healthcare), using 1mm thick, 24-cm, 13% polyacrylamide gels, and electrophoresis were carried out at a constant current (45 min at 2 W/gel, then at 1 W/gel for ~16 h at 25 °C). 2D gels were scanned on a Typhoon Trio Imager according to manufacturer's protocol (GE Healthcare) at 100 μ m resolution with excitation wavelength of 532 nm with a long pass emission wavelength of 570 nm and 488 nm with 520 nm emission for Z-Dye, and 633nm with long pass emission of 670 nm for Cy5. The number of visible protein spots was counted for direct

comparison; fluorescent measurements and image processing was done using ImageJ (Rasband et al. 2012).

Metabolite Extraction

A survey of the *S. solfataricus* metabolome was done by extracting metabolites using 60% aqueous (v/v) EtOH, 50% aqueous (v/v) MeOH or MeOH/Chloroform (1 mM EDTA) solvents in an effort to ensure global coverage. Chloroform methanol extraction was a modification from (Canelas et al. 2009). Briefly, cell pellets were resuspended in 1.5ml of cold MeOH and transferred to glass tubes, after which 1.5 ml of cold chloroform was added. All samples were then shaken at 0°C for 2 hours using a custom made orbital shaker. The samples were then centrifuged (2000 x g, 15 min, -9°C). Upper MeOH/Chloroform layer was removed and 50% aqueous MeOH (v/v) was added to cellular debris and vortexed for 30s. After centrifugation the upper phase was pooled with the first extracts. Proteins were precipitated using 5:1 dilution with -80°C acetone and centrifuged. Upper phase was collected and subsequently dried and resuspended in 50% MeOH. MeOH/H₂O, EtOH/H₂O extraction was a modification from (31). Briefly each sample was resuspended in 1.5 ml 60% aqueous EtOH or 50% MeOH (v/v) respectively, 1 mM EDTA, vortexed for 30s, and sample was frozen in liquid nitrogen and thawed to room temperature three times. Samples were subsequently sonicated 60% duty cycle maximum power level for 5 min on ice. The sample was incubated 1 hr at -20°C followed by a 15 min centrifugation at 13000 x g and supernatant was pooled. Cell debris was subsequently washed with 1 ml extraction solvent, vortexed for 30s and centrifuged. Proteins were precipitated using 5:1 dilution with -80C acetone and

centrifuged. Upper phase was collected and subsequently dried and resuspended in 50% MeOH. *Ignicoccus hospitalis* and *E. coli* metabolites were extracted using identical 50% MeOH method.

LCMS Based Metabolome Analysis

Analyses were performed using a 1290 UPLC coupled to a 6538 UHD Accurate-Mass Q-TOF (Agilent Technologies). The system was operated in positive and negative electrospray ionization modes. Vials containing extracted metabolites and standard mixtures were kept at -80°C prior to LCMS analysis. Metabolites were separated using a reverse-phase Kinetix 1.7µm C18, 100A, 150mm x 2.1mm (Reverse Phase), or a Cogent Diamond Hydride, 150mm, (HILIC). In positive ionization mode, A= 0.1% formic acid in water B=0.1% formic acid in acetonitrile. For negative ionization mode two pairs of solvent systems were used, A1= 0.1% formic acid in water B1=0.1% formic acid in acetonitrile, A2= 1mM Ammonium Fluoride in water B2= acetonitrile. The C18 linear gradient was 100% A (0 - 2 min) – 100%B with an injection volume of 15 µL. HILIC linear gradient was 5% A (0 – 2 min) – 50% A with an injection volume of 2 ul. ESI conditions were modified from (25) gas temperature 250°C, drying gas of 12 L/min, nebulizer 30psig, fragmentor 120V, and skimmer 65V. The instrument was set to acquire over the m/z range 50-1000 with an acquisition of 1 spectra/second. MSMS fragmentation was performed in targeted mode, and the instrument was set to acquire over the m/z range 50-1000, with default isowidth, collision energy was 10-20V. Data was visualized using MassHunter software package (Agilent Technologies) and data processing was done using XCMS software and XCMS online, small molecule

characterizations were done using the Metlin database (Tautenhahn et al. 2008, Smith et al. 2005, Patti et al. 2013, Tautenhahn et al. 2012).

GSH Labeling and GSSG Reduction

A method for labeling reduced glutathione with N-(2-aminoethyl) maleimide was developed using a GSH standard (Sigma) in aqueous methanol (50%, pH 6.5). Using HILIC chromatography described above, a characteristic mass and retention time shift allowed semi-quantitative analysis of labeling to be performed. To test the reaction conditions that were developed, 100 nmol GSH was mixed with 35, 70, and 140 nmol of N-(2-aminoethyl) maleimide in which a linear reduction of GSH and linear accumulation of product was observed (Supplementary Fig. 4A). Metabolite extracts from *S. solfataricus* and *E. coli* were then spiked with 100 nmol N-(2-aminoethyl) maleimide in triplicate and analyzed with XCMS software. Using HILIC methods to visualize GSSG reduction, GSSG standard was spiked into samples (50% MeOH) and reduced, using 10 nmol, 100 nmol and 1 μ mol dithioreitol (DTT). Metabolite extracts from *S. solfataricus* were then spiked with 2 μ mol DTT in triplicate and analyzed with XCMS software to identify potentially novel small molecule thiols.

Determination of NADPH-ratio

The intracellular ratio in *S. solfataricus* and *E. coli* (BL21-D3) cells was determined as described by the manufacture of the NADP/NADPH Quantitation Kit, MAK038-1KT (Sigma Inc). Cells were grown to mid-log and stationary phase, with NADP(H) extracted immediately before measurement.

Results

The *S. solfataricus* Proteome Has Less Cysteine and More Disulfide Bonds

Genomic analysis revealed that the proteins of *S. solfataricus* have a lower percentage of cysteine than proteins from bacteria (*Escherichia coli*) and eukarya (*Saccharomyces cerevisiae*). Across the proteome of *S. solfataricus* cysteine residues make up only 0.6% of the amino acid composition, compared with 1.2% in *E. coli* and 1.3% in *S. cerevisiae*. In addition, over a third (36.5%) of the proteins from *S. solfataricus* do not contain cysteine, a distinct difference from *E. coli* (15.2%) and *S. cerevisiae* (9.3%).

To determine if a significant fraction of cytosolic proteins in *S. solfataricus* contained oxidized (PSSP), rather than reduced cysteine (PSH), a fluorescent thiol reactive maleimide probe was used to label intracellular proteins. Free cysteine thiols in control and reduced protein extracts were visualized by running the labeled proteins on 2 dimensional gels. The number and intensity of protein spots was substantially increased after reduction (Fig. 1). The increase in reduced thiols after reduction was confirmed by analyzing scanned images of the gels run in triplicate, which showed that average fluorescence intensity increased from 1126 ± 77 relative units in the control, to 1797 ± 282 after reduction. This supports predictions based on sequence analysis, that the intracellular proteins of *S. solfataricus* have an over representation of paired cysteines (Mallick et al. 2002, Beeby et al. 2005, Jorda et al. 2011).

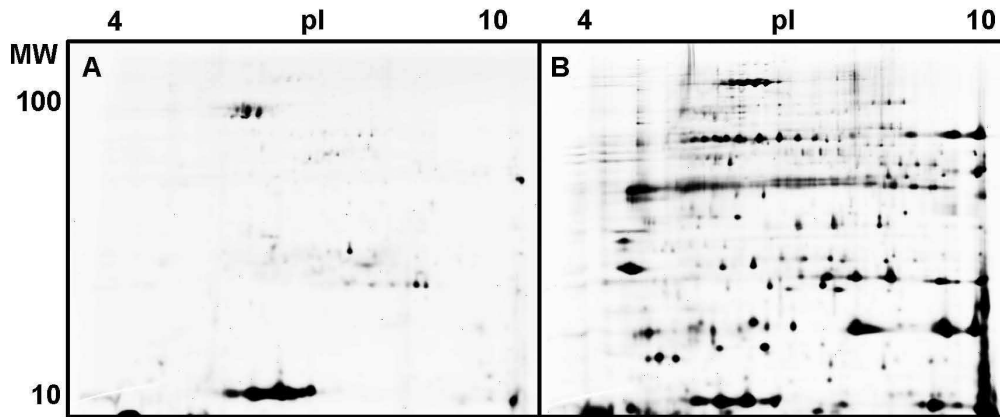


Figure 3.1 Cysteine Labeling of Cytoplasmic Proteins of *S. solfataricus* Before and After Reduction of Disulfide Bonds. Comparison of reduced thiol content in total protein sample using a fluorescent Z dye-coupled maleimide probe in the native state (A) and after reduction with tributylphosphine (B). The number of proteins labeled and total fluorescence intensity increased after reduction of disulfide bonds.

The Metabolome of *S. solfataricus* Favors Oxidized Thiol

Results from the analysis of the proteome provide evidence that the intracellular thiol environment of *S. solfataricus* is unlike that reported in all eukaryotes and all but a few prokaryotes. A detailed characterization of the metabolome was undertaken to identify any small molecules involved with maintaining the unusually oxidized PSSP, and find a second indicator of intracellular redox potential. To maximize the metabolic coverage, ultra performance liquid chromatography based mass spectrometry (UPLC/MS) was used to analyze a series of metabolite extractions using multiple solvents (Supplemental Fig. 1). In batch culture, cellular metabolite expression is dependent upon the phase of an organism's growth (Supplemental Fig. 2). Therefore, metabolites from *S. solfataricus* were extracted from cultures grown at mid-log and stationary phases as well as after 30 minutes of oxidative stress with 30 μ M H_2O_2 .

Untargeted metabolite profiles were compared using both reverse phase and hydrophilic interaction chromatography, to ensure deep coverage (Fig. 2A, B). Analysis of

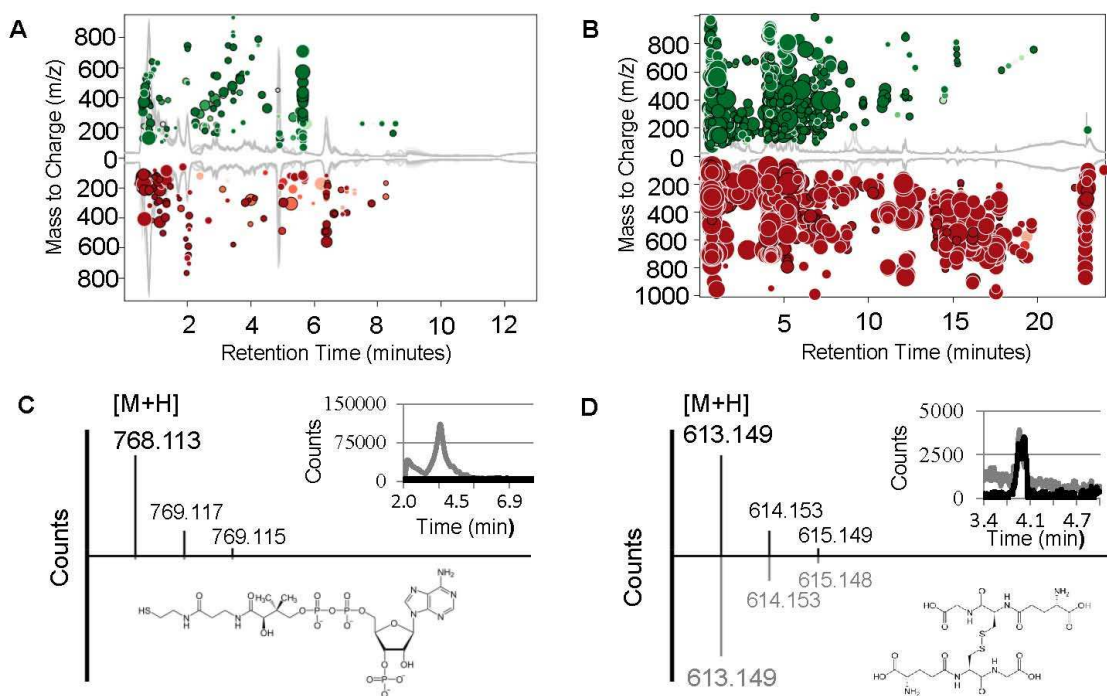


Figure 3.2 Metabolic Profiling and Characterization of Thiol Content. (A) Full chromatographic reconstruction of all mass spectral features detected and associated retention times using LCMS was able to discriminate 615 and 556 molecular features in reverse phase and HILIC separations, respectively. (B) Lack of CoA in metabolite extracts was confirmed by spike-in experiments which show a clear mass spectrum of CoA standard. Inset, CoA standard(gray), and absence in metabolite extracts (black). (C) Presence of oxidized glutathione (GSSG) in *S. solfataricus*. Mass spectra of GSSG from *S. solfataricus* (above) compared to GSSG standard (below). Inset, retention time comparison between GSSG standard (gray) and GSSG native to *S. solfataricus* (black), structure of GSSG shown for reference.

metabolite profiles after 30 minutes of oxidative stress with 30 μ M H_2O_2 , allowed visualization of 2065 unique small molecule masses (fold change ≥ 2 , p-value ≤ 0.01); none of which matched the mass of common reduced small molecule thiol, such as

Cysteine, GSH, or Coenzyme A (CoA) (Fig. 2C). CoA was previously suggested as a redox mediator for *S. solfataricus* and related thermophiles (Hummel et al. 2005).

While reduced GSH was not detectable, GSSG was present in low abundance (~12.5 μ M) (Fig. 2D). To ensure the lack of appreciable GSH was not caused by the metabolite extraction method, *E. coli* metabolites were extracted from stationary phase using an identical method. GSH was highly abundant. Based upon cell mass and LCMS spectral intensity, the amount of GSSG detected in *S. solfataricus* averaged only 0.62% (p-value < 0.05) that found in *E. coli*. Previous reports failed to find glutathione in *S. solfataricus* (Hummel et al. 2005), suggesting our success may be attributed to the superior sensitivity of UPLC/MS. In most prokaryotic and eukaryotic cells, GSH is typically found in greater abundance than GSSG during normal healthy growth (Sevier et al. 2002). To further validate this unique redox poise, metabolites from the anaerobic thermophilic archaea *Ignicoccus hospitalis* were extracted and similar low levels of GSSG were detected, again in the absence of measurable GSH (Supplemental Fig. 3).

Detection of Small Molecules Using Chemical Biology

The failure to detect a common reduced small molecule thiol in *S. solfataricus* did not rule out the existence of a novel form, such as that observed in eubacteria and lower temperature thermophilic archaea (Allocati et al. 2012, Malki et al. 2009). Therefore, we adopted a chemical labeling approach to search for thiol containing small molecules. The method allows for direct labeling of thiol in cellular extracts using a maleimide reactive group. To test the protocol, reaction of GSH with N-(2Aminoethyl)maleimide was

monitored using UPLC/MS. Extracted ion chromatograms (EIC) showed a decrease in substrate and reciprocal accumulation of product (Supplemental Fig. 4A, B), demonstrating that the assay performed as expected. The maleimide probe was then spiked into metabolite extracts from both *S. solfataricus* and *E. coli*. Targeted analysis of *E. coli* extracts showed native GSH was readily labeled (18-fold reduction, p-value < 0.05) along with several other putatively identified thiols (Table S1), but in the *S. solfataricus* samples, no native reduced thiol reacted with the probe. GSH and N-(2-aminoethyl) maleimide were iteratively spiked into *S. solfataricus* metabolite extract, to validate that there was no unknown interfering small molecule or matrix effects in the extracts. Clear labeling of GSH standard was observed, confirming that if GSH was present at $\geq 25 \mu\text{M}$, it would have been successfully labeled (Supplemental Fig. 5). Glutathione can exist at mM concentrations in eukaryotic mesophiles (Smirnova et al. 2005).

To probe the metabolome of *S. solfataricus* for the presence of other oxidized disulfides, a method for reducing oxidized thiols in metabolite extracts was developed. Using this approach GSSG was reduced to GSH by dithiothreitol (DTT) in a dose dependent manner. Differential analysis of the metabolome did not reveal other affected small molecules, strongly supporting a lack of other disulfides. A concentration threshold for the reaction was observed, below which oxidized disulfides could not be reduced. The efficiency of disulfide reduction decayed as the available GSSG concentration decreased (Supplemental Fig. 4C). The inability of DTT to reduce all

detectable GSSG in *S. solfataricus*, explains why it would not have been detected previously (Hummel et al. 2005).

NADPH/NADP⁺ Ratio

NADPH provides reducing equivalents in anabolic pathways and facilitates regeneration of GSH through glutathione reductase (Pai et al. 1983). The ratio of NADPH to NADP⁺ in *S. solfataricus* was determined using a standard colorimetric assay (Sigma Inc.). The results showed that significantly higher levels of oxidized NADP⁺ were present (Supplemental Fig. 6). While the ratio of oxidized to reduced equivalents of glutathione and NADPH are not directly connected, this provides further evidence that related metabolite pools of *S. solfataricus* maintain a more oxidized potential.

Genomic Justification for GSSG

The experimental evidence indicated GSSG exists in *S. solfataricus*; however, substantial pathway annotation was missing. With the exception of gamma-glutamyl transpeptidase (Sso3216) and disulfide protein oxidoreductase (SsDPO) (Pedone et al. 2006), requisite enzymes for the synthesis of GSH were not known to be present. Updated genome annotation performed by SulfoSYS (Albers et al. 2009), revealed an archaeal cluster of orthologous genes (COG) to glutathione synthase, lactoyl-glutathione lyase, and glutaredoxin-like proteins. Discovery of GSSG prompted us to investigate these gene clusters further, with the goal of completing the synthetic pathway. While sequence based comparisons often fail when there is high divergence, even low identity can provide evidence of homologous structures and activities (Heinemann et al.

2011). To elucidate the glutathione pathway, genes for putative enzymes were analyzed using machine learning approaches (Dlakic et al. 2009, Shah et al. 2008). This resulted in the identification of three genes with significant similarity to those of glutathione synthase, a protein disulfide oxidoreductase with experimentally validated GSH-thiol transferase activity (Pedone et al. 2006), and confirmation of a newly annotated gamma-glutamyltransferase (Table 3.1).

Table 3.1 Summary of *S. solfataricus* protein with similarity to glutathione synthesis and metabolism

Gene	arCOG Annotation	arCOG #	P-value	8
SSO0159	Glutathione synthase/glutaminyl transferase/alpha-L-glutamate ligase	arCOG01589	2.8E-45	
SSO0645	Glutathione synthase/glutaminyl transferase/alpha-L-glutamate ligase	arCOG01589	6.2E-39	
SSO2346	Glutathione synthase/glutaminyl transferase/alpha-L-glutamate ligase	arCOG01589	3.3E-37	
SSO1223	Catechol 2,3-dioxygenase or other lactoylglutathione lyase family enzyme	arCOG06037	0	
SSO2054	Catechol 2,3-dioxygenase or other lactoylglutathione lyase family enzyme	arCOG06037	0	
SSO2426	Lactoylglutathione lyase or related enzyme	arCOG02706	5E-24	
SSO0192*	Thiol-disulfide isomerase or thioredoxin (SsPDO)	arCOG01218	ref [8]	
SSO0314	Predicted thioredoxin/glutaredoxin	arCOG04235	0	
SSO3216*	Gamma-glutamyltransferase	arCOG04053	Annotated	

SSO0192* has been biochemically characterized and displays activity with GSSG and GSH; SSO3216* is annotated.

Discussion

This report provides evidence from multiple systems that the ratio of oxidized to reduced thiol in *S. solfataricus* is substantially shifted toward the oxidized form. This cellular state does not appear to be maintained by novel pathways or enzymes, but through adjustment of the canonical redox system. Global analysis of intracellular protein disulfide shows there is a substantial increase in the number of fluorescently labeled proteins after reduction, an observation which contrasts most bacterial and eukaryotic systems (Sevier et al. 2009). Previous investigations of thiol content in thermophilic organisms reported an absence of glutathione or alternatives, such as γ -glutamylcysteine found in archaea from cooler environments (Allocati et al. 2012, Malki et al. 2009). We confirmed the absence of small molecule thiol by chemically labeling reduced thiol in *S. solfataricus* metabolite extracts. However, our large scale metabolomic survey detected glutathione in an oxidized form, GSSG. This result prompted a careful investigation of the *S. solfataricus* genome, in which genes with significant similarity to those involved in glutathione synthesis and regulation were validated. Enzyme assays on protein disulfide oxidoreductase (SsPDO) from *S. solfataricus* show that glutathione (GSH) dependent thiol-transferase activity is suppressed (Pedone et al. 2006). This would help establish/maintain oxidized protein disulfides, because the GSH/GSSG ratio is linked to PSH/PSSP ratio through direct and indirect redox transfer reactions. Connectivity of chemical potential between small molecules and proteins allows cells to take advantage of substantial redox buffering

capacity of the protein thiol pool (Hansen et al. 2009, Cumming et al. 2004, Chakravarthi et al. 2006, Pastore et al. 2003).

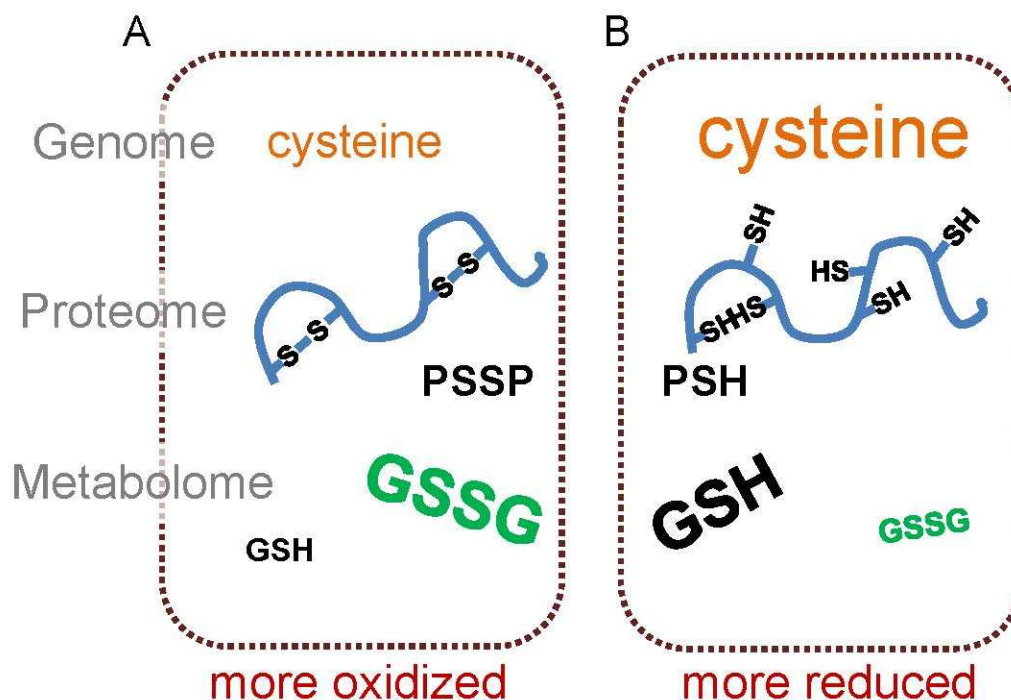


Figure 3.3 Results of Genomic, Proteomic and Metabolomic Investigation Comparing Thiol Content. Thermophilic *S. solfataricus* compared to mesophilic organisms *Escherichia coli* and *Saccharomyces cerevisiae*. (Bottom) Placement of *S. solfataricus*, *E. coli* and *S. cerevisiae* in rRNA based phylogenetic tree (after 31). (A) Thiol and redox conditions typically observed in mesophilic organisms. The ratio of oxidized (GSSG) to reduced glutathione (GSH) is much higher in *S. solfataricus* (B). Genomic analysis indicates a reduced use of the amino acid cysteine in the proteins of *S. solfataricus*. Proteomic investigation indicates that cysteines (PSH) in *S. solfataricus* occur primarily as protein disulfides (PSSP). Combined these results promote a system-wide model for redox homeostasis in thermophilic organisms.

Historically, it was suggested proteins in archaea were low in cysteine residues and disulfide bonds (Ladenstein et al. 2008), but more recently evidence shows significant disulfide bonding in thermophilic archaea (Mallick et al. 2002, Beeby et al.

2005, Jorda et al. 2011). In the case of *S. solfataricus*, the amount of free thiol in the proteome and metabolome is lower, and the abundance of intracellular protein disulfides are higher than bacterial or eukaryotic model organisms, prompting us to put forth a new thiol-redox strategy (Fig. 3).

Many of the proteins from *S. solfataricus* show an absence of cysteine (36.5%), but this is not surprising considering studies of thermophilic viruses which show the membrane proteins often lack any cysteine (Menon et al. 2008). Combined with previous results it seems likely thermophiles utilize a more oxidizing intracellular environment to take advantage of the structural stability imparted by disulfides. To achieve this, thermophilic organisms need to maintain an intracellular environment in which protein disulfides persist. In this case, a glutathione based thiol pool may not be appropriate as a redox buffer. It has been previously suggested that coenzyme A could take on this role, but a lack of detectable levels in this study (Fig. 2C) and others (Hummel et al. 2005) suggests this is unlikely. An alternative explanation is that *S. solfataricus* has reduced the role of thiol in redox chemistry, freeing disulfides to be used extensively for thermostability of intracellular proteins. Consistent with this model is the fact that while disulfide number is increased, the overall percentage of cysteine residues is reduced because of the potential to form unwanted crosslinks. Previous investigations of the cellular response to oxidative stress lend further support to an alternative cytoplasmic potential. We found that during periods of oxidative stress the gene encoding gamma-glutamyltransferase (Sso3216) is up-regulated, and the putative protein glutathione

synthase (Sso0159) is down-regulated in *S. solfataricus*, which would trigger a depletion of the GSH pool (Maaty et al 2009).

By using an integrated analysis of metabolites and proteins, a unified model of thermophile redox physiology is put-forth. A point that remains to be addressed is the role of glutathione in *S. solfataricus* and whether this represents an ancestral cellular state from a time before reactive oxygen species altered the landscape of biology (Schwartzman et al. 2005). A substantial thiol pool exists in mesophilic organisms from all branches of life; even mesophilic archaea utilize a glutathione-like thiol pool. Based on the placement of thermophilic archaea close to the last universal common ancestor in rRNA phylogenies, we suggest that thiol-based systems now used for maintaining a reducing intracellular environment are derived from metabolic pathways originally tasked with keeping protein disulfide bonds intact at high temperature.

Acknowledgements

The authors thank Edward Schmidt for helpful comments on the manuscript, Susan K. Brumfield for SEM images in Graphic Abstract, Merrill Oakley and Michelle Tigges for work with metabolite extractions, and Rebecca Sorenson and Sophia Froelich for help with bioinformatics. This work was supported by National Science Foundation, MCB0646499 and MCB102248. Mass Spectrometry, Proteomics and Metabolomics core facility supported by the Murdock Charitable Trust, INBRE MT Grant No. P20 RR-16455-08, NIH Grant Nos. P20 RR-020185 and P20 RR-024237 from the COBRE Program of the National Center for Research Resources.

References

- Albers SV, Birkeland N, Driessen AJ, Gertig S, Zaparty M, SulfoSYS (Sulfolobus Systems Biology): towards a silicon cell model for the central carbohydrate metabolism of the archaeon *Sulfolobus solfataricus* under temperature variation, *Biochem. Soc. Trans.* 37 (2009) 58–64.
- Allocati N, Federici L, Masulli M, Di Ilio C, Distribution of glutathione transferases in Gram-positive bacteria and Archaea, *Biochimie.* 94 (2012) 588–596.
- Beeby M, O'Connor BD, Ryttersgaard C, Boutz D, Yeates TO, The genomics of disulfide bonding and protein stabilization in thermophiles, *PLoS Biol.* 3 (2005) e309.
- Canelas AB, Pierick A, Ras C, Siefar RM, Heijnen JJ, Quantitative evaluation of intracellular metabolite extraction techniques for yeast metabolomics, *Anal. Chem.* 81 (2009) 7379–89.
- Chakravarthi S, Jessop CE, Bulleid NJ, (2006) The role of glutathione in disulphide bond formation and endoplasmic-reticulum-generated oxidative stress, *EMBO Rep.* 7 (2006) 271–275.
- Cumming RC, Andon NL, Haynes PA, Minkyu P, Schubert D, Protein disulfide bond formation in the cytoplasm during oxidative stress, *J. Biol. Chem.* 279 (2004) 21749–21758.
- Derman J, Beckwith J, (1991) *Escherichia coli* alkaline phosphatase fails to acquire disulfide bonds when retained in the cytoplasm, *J. Bacteriol.* 173, 7719–7722.
- Dlakić M, HHsvm: fast and accurate classification of profile-profile matches identified by HHsearch, *Bioinformatics.* 25 (2009) 3071–3076.
- Fahey RC, (2001) Novel thiols of prokaryotes, *Annu. Rev. Microbiol.* 55, 333–356.
- Smirnova GV, Oktyabrsky ON, (2005) Glutathione in bacteria. *Biochemistry, Biokhimii* a. 70, 1199–1211.
- Fratelli M, (2005) Gene expression profiling reveals a signaling role of glutathione in redox regulation, *Proc. Natl. Acad. Sci. USA.* 102, 13998–14003.
- Grieco PA, Young MJ, Bothner B, (2012) Global analysis of viral infection in an archaeal model system, *Front. Microbiol.* 3, 411.
- Hansen RE, Roth D, Winther JR, Quantifying the global cellular thiol-disulfide status, *Proc. Natl. Acad. Sci. USA.* 106 (2009) 422–427.

Heinemann J, Hamerly T, Maaty WS, Movahed N, Steffens JD, Reeves BD, Hilmer, JK, Therien J, Grieco PA, Peters JW, Bothner B, (2014) Expanding the paradigm of thiol redox in the thermophilic root of life. *Biochimica et Biophysica Acta* 1840, 80-85.

Jorda J, Yeates TO, (2011) Widespread disulfide bonding in proteins from thermophilic archaea, *Archaea*. 409156.

Hummel CS, Lancaster KM, Edward J, Crane EJ, (2005) Determination of coenzyme A levels in *Pyrococcus furiosus* and other Archaea: implications for a general role for coenzyme A in thermophiles, *FEMS Microbiol. Lett.* 252, 229–234.

Ladenstein L, Ren B, (2008) Reconsideration of an early dogma, saying "there is no evidence for disulfide bonds in proteins from archaea", *Extremophiles*. 12, 29-38.

Lamble HJ, Theodossis A, Milburn CC, Taylor GL, Danson MJ, (2005) Promiscuity in the part-phosphorylative Entner-Doudoroff pathway of the archaeon *Sulfolobus solfataricus*, *FEBS Lett.* 579, 6865–6869.

Larson ET, Eilers B, Menon S, Reiter D, Lawrence CM, (2007) A winged-helix protein from *Sulfolobus* turreted icosahedral virus points toward stabilizing disulfide bonds in the intracellular proteins of a hyperthermophilic virus, *Virology*. 368, 249–261.

Lawrence CM, Menon S, Eilers BJ, Bothner B, R. Khayat R, Douglas T, Young MJ, (2009) Structural and functional studies of archaeal viruses, *J. Biol. Chem.* 284, 12599-603.

Maaty WM, Selvig K, Ryder S, Tarlykov P, Hilmer JK, Heinemann J, Steffens J, Snyder JC, Ortmann AC, Movahed N, Spicka K, Chetia L, Grieco PA, Dratz EA, Douglas T, Young MJ, Bothner B. (2012) Proteomic analysis of *Sulfolobus solfataricus* during *sulfolobus* turreted icosahedral infection, *Journal of proteome research*. 11, 1420-32.

Maaty WS, Steffens J, Heinemann J, Ortmann AC, Reeves BD, Biswas SK, Dratz EA, Grieco PA, Young MJ, Bothner B, (2012) Global analysis of viral infection in an archaeal model system, *Front. Microbiol.* 3, 411.

Malki L, Yanku M, Borovok Y, Cohen G, Aharonowitz Y, (2009) Identification and characterization of *gshA*, a gene encoding the glutamate-cysteine ligase in the halophilic archaeon *Haloferax volcanii*, *J. Bacteriol.* 191, 5196–5204.

Menon SK, Maaty WS, Corn GJ, Bothner B, Lawrence CM, (2008) Cysteine usage in *Sulfolobus* spindle-shaped virus 1 and extension to hyperthermophilic viruses in general, *Virology*. 376, 270–278.

Nunn CE, Johnsen U, Schonheit P, Fuhrer T, Danson MJ, Metabolism of pentose sugars in the hyperthermophilic archaea *Sulfolobus solfataricus* and *Sulfolobus acidocaldarius*, *J. Biol. Chem.* 285 (2010) 33701–33709.

Pai EF, Schulz GE, The catalytic mechanism of glutathione reductase as derived from x-ray diffraction analyses of reaction intermediates, *J. Biol. Chem.* 258 (1983) 1752–57.

Pastore A, Federici G, Bertini E, Piemonte F, (2003) Analysis of glutathione: implication in redox and detoxification, *Clin. Chim. Acta.* 333, 19–39.

Patti GJ, Tautenhahn R, Siuzdak G, (2012). Meta-analysis of untargeted metabolomic data from multiple profiling experiments. *Nature Protocols*, 7(3), 508–16.

Patti GJ, Tautenhahn R, Rinehart D, Cho K, Shriver LP, Manchester M, Nikolskiy I, Johnson CH, Mahieu NG, Siuzdak G, (2013) A View from Above: Cloud Plots to Visualize Global Metabolomic Data, *Anal. Chem.* 85, 798–04.

Pedone E, Limauro D, D’Alterio R, Rossi M, Bartolucci S, (2006) Characterization of a multifunctional protein disulfide oxidoreductase from *Sulfolobus solfataricus*, *FEBS J.* 273, 5407–5420.

Rasband WS, ImageJ. U.S. National Institutes of Health, Bethesda, Maryland, USA. (2012) Available at: <http://rsbweb.nih.gov/ij/>.

Schwartzman D, Lineweaver CH, Biospheric Self-Organization A Hyperthermophilic Origin of Life, in: A. Kleidon, R.D. Lorenz, Non-equilibrium Thermodynamics and the production of Entropy. In Non-equilibrium Thermodynamics and the production of Entropy, Springer-Verlag Berlin Heidelberg, Germany, 2005, pp. 207–221.

Sevier CS, Kaiser C (2002) Formation and transfer of disulphide bonds in living cells, *National review molecular cell biology.* 3, 836–47.

Shah AR, Oehmen CS, Webb-Robertson B, (2008) SVM-HUSTLE--an iterative semi-supervised machine learning approach for pairwise protein remote homology detection, *Bioinformatics.* 24, 783–90.

She Q, Singh RK, Confalonieri F, Zivanovic, J. van der Oost J, (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2, *Proc. Natl. Acad. Sci. USA.* 98, 7835–7840.

Smith C, O'Maille G, Want EJ, Qin C, Siuzdak G, METLIN: a metabolite mass spectral database, *Therapeutic drug monitoring.* 27 (2005) 747–51.

Tautenhahn R, Böttcher C, Neumann S, (2008) Highly sensitive feature detection for high resolution LC/MS, *BMC Bioinformatics*. 9, 504.

Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G, XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data, *Anal. Chem.* 84 (2012) 5035-39.

Weidenheft B, Mosolf J, Willits D, Yeager M, Dryden KA, Young M, Douglas T, (2005). An archaeal antioxidant: characterization of a Dps-like protein from *Sulfolobus solfataricus*, *Proceedings of the National Academy of Science of the United States of America*. 102(30), 10551-6.

Yanes O, Tautenhahn GJ, Patti G, Siuzdak G, (2011). Expanding coverage of the metabolome for global metabolite profiling, *Analytical chemistry*. 83, 2152–61.

CHAPTER FOUR

APPLICATION OF SUPPORT VECTOR MACHINES TO METABOLOMICS
EXPERIMENTS WITH LIMITED REPLICATES.Contribution of Authors and Co-Authors

Manuscript in Chapter 4

Author: Joshua Heinemann

Contributions: Culturing of *S. solfataricus*; Extracted metabolites; Ran untargeted metabolomics on MS; Formatted data; Designed experiments; Wrote manuscript.

Co-Author: Aurélien Mazurie

Contributions: Designed processing pipeline; Used cloud to run over 1 million jobs; Wrote manuscript.

Co-Author: Monika Tokmina-Lukaszewska

Contributions: Extracted metabolites; Ran untargeted metabolomics on MS; Wrote manuscript.

Co-Author: Greg J. Beilman

Contribution: Provided valuable insight; Procured funding; Provided *Sus scrofa* samples.

Co-Author: Brian Bothner

Contributions: Provided valuable insight; Wrote manuscript; Procured funding.

Manuscript Information Page

Authors: Joshua Heinemann, Aurélien Mazurie, Monika Tokmina-Lukaszewska, Greg J. Beilman and Brian Bothner

Journal: Metabolomics

Status of the manuscript: Accepted

☐ Prepared for submission to a peer-reviewed journal

☐ Officially submitted to a peer-reviewed journal

☒ Accepted by a peer-reviewed journal

☐ Published in a peer-reviewed journal

Publisher: Springer

Issue in which manuscript appears: To be determined

APPLICATION OF SUPPORT VECTOR MACHINES TO METABOLOMICS
DATASETS WITH LIMITED REPLICATES

Joshua Heinemann¹, Aurélien Mazurie², Monika Tokmina-Lukaszewska¹, Greg J.
Beilman⁴ and Brian Bothner^{1,3}

Department of Chemistry and Biochemistry¹, Bioinformatics Core Facility², Mass
Spectrometry Core Facility³ Montana State University, Bozeman, MT 59717, Department
of Surgery⁴, University of Minnesota, Minneapolis, MN 55433

Abstract

Identifying biomarkers that are indicative of a phenotypic state is difficult because of the amount of natural variability which exists in any population. While there are many different algorithms to select biomarkers, previous investigation shows the sensitivity and flexibility of support vector machines (SVM) make them an attractive candidate. Here we evaluate the ability of support vector machine recursive feature elimination (SVM-RFE) to identify potential metabolic biomarkers in liquid chromatography mass spectrometry (LCMS) untargeted metabolite datasets. Two separate experiments are considered, a low variance (low biological noise) prokaryotic stress experiment, and a high variance (high biological noise) mammalian stress experiment. For each experiment, the phenotypic response to stress is metabolically characterized. SVM-based classification and metabolite ranking is undertaken using a systematically reduced number of biological replicates to evaluate the impact of sample size on biomarker reproducibility and

robustness. Our results indicate the highest ranked 1% of metabolites, the most predictive of the physiological state, were identified by SVM-RFE even when the number of training examples was small (≥ 3) and the coefficient of variation (CV) was high (> 0.5). An accuracy analysis shows filtering with recursive feature elimination measurably improves SVM classification accuracy, an effect that is pronounced when the number of training examples is small. These results indicate that SVM-RFE can be successful at biomarker identification even in challenging scenarios where the training examples are noisy and the number of biological replicates is low.

Introduction

Cellular life utilizes a wide range of metabolic phenotypes to survive in different environments. These phenotypes can be discriminated using liquid chromatography mass spectrometry (LCMS) based untargeted metabolite profiles, which are sensitive to a wide variety of environmental influences including both disease and stress (Patti et al. 2012, Herder et al. 2011, Serkova et al. 2011). Ultimately, biologists would like to identify biomarkers that link to a disease, stress, or environment, however, normal biological variation in metabolic response can make this difficult. Support Vector Machines (SVMs) are machine learning classifiers designed to mitigate the effects of noisy data and have recently been employed for biomarker discovery using SVM-recursive feature elimination (SVM-RFE) (Lin et al. 2011, Guyon et al. 2002). Results indicate that SVM-RFE works well when the number of training samples is small and the number of features

is large (Lin et al. 2011, Guyon et al. 2002), however, further investigation into the effect of total sample number on experimental results is warranted.

When used for biomarker detection SVM-RFE is employed as a backwards elimination algorithm which ranks metabolites by their ability to accurately discriminate two phenotypes. In this way SVM-RFE creates a ranking system, where the features best at discriminating phenotype are ranked higher than less informative features. This ranking system is convenient for biomarker discovery as it ranks based upon predictive accuracy. Unfortunately, it is difficult to determine if a sample population is of sufficient size to identify accurate metabolic biomarkers for a larger population. While SVM has been highly successful when applied to datasets with large numbers of training samples: e.g., $n=78$, 42, 38, 32 (Duan et al. 2005), $n=34$ (Bertini et al. 2009), $n=72$ (Guan et al. 2009), and $n=59$ (Mahadevan et al. 2008), we are unaware of studies on biological systems where the effect of the number of training samples is tested.

The number of biological samples accessible to researchers is often limited by the cost and feasibility of collecting data. . Here we provide evidence that biomarkers selected by SVM-RFE are robust even in cases where the number of training examples is low. We find the sample size has little influence on the top 1% of ranked metabolites when SVM-RFE is trained with three or more samples representative of each phenotype. The highest ranked biomarkers remain predictive when applied to larger populations. Even in an unfavorable experimental scenario, where the biological noise is high and the number of training samples is low, SVM-RFE can identify robust biomarkers.

Materials and Methods

Culturing of *Sulfolobus solfataricus*

S. solfataricus was grown aerobically in liquid DSMZ media 182 (22.78 mM KH_2PO_4 , 18.90 mM $(\text{NH}_4)_2\text{SO}_4$, 0.81 mM MgSO_4 , 1.7 mM CaCl_2 , 0.2% Yeast Extract) pH adjusted to 2.8 with 6 M H_2SO_4 . Batch cultures of *S. solfataricus* were grown in long neck Erlenmeyer flasks at 80°C. One liter of media was inoculated with 10 ml of log phase (OD_{650} 0.33) culture and divided evenly between ten, 1-liter long neck flasks. At (OD_{650} , 0.35), 50 mL of each 0.25 liter culture was removed and placed in a 50 ml falcon tube as baseline controls. H_2O_2 was then administered to a final concentration of 30 M (n=10). An additional 50 mL aliquot was collected from each culture at 30 min post H_2O_2 inoculation and used for metabolite extraction using a method modified from (Maaty et al., 2009).

Metabolite Extraction from *S. solfataricus*

Metabolite extraction was done using 50% aqueous (v/v) MeOH, a modification from (Yanes, et al., 2011). Briefly each sample was resuspended in 50% MeOH (v/v) respectively, 1 mM EDTA, vortexed for 30s, and sample was frozen in liquid nitrogen and thawed to room temperature three times. Samples were subsequently sonicated 60% duty cycle maximum power level for 5 min on ice. The sample was incubated 1 hr at -20°C followed by a 15 min centrifugation at 13,000 x g and supernatant was pooled. Cell debris was washed with 1 ml extraction solvent, vortexed for 30s and centrifuged.

Proteins were precipitated using 5:1 dilution with -80°C acetone and centrifuged. Upper phase was collected and subsequently dried and resuspended in 50% MeOH for injection.

S. scrofa Urine Collection and Hemorrhagic Shock Protocol

An already established model of porcine hemorrhagic shock was described (Luszczek, et al. 2011; Scribner, et al. 2010; Mulier, et al. 2005). Briefly, hemorrhagic shock was induced on 8 male Yorkshire pigs (Manthei Hog Farm, LLC, Elk River, MN) while anesthetized; a blunt percussive injury to the chest was created using a captive bolt device followed by withdrawal of blood (35%) from the inferior vena cava. Final trauma was induced by two crush injuries created in the liver parenchyma. To secure urine sampling, a Foley catheter was placed in urinary bladder during the surgery. Baseline urine samples were collected prior to induction of shock; urine shock samples were collected after 2 hrs. Urine was kept at -80 °C till used for analysis.

Metabolite Purification from S. scrofa Urine

For *Sus scrofa* samples, raw urine samples were spun at 10,000 x g for 15min and then supernatants (200 L) were carefully placed in nanosep spin device (Pall Corporation) equipped with 3.5 kDa molecular weight cutoff membranes. Samples were diluted with 200 L of water and spun at 13,000 x g for 15min. Afterwards membrane was treated twice with 100ul of 10% MeOH solution. Eventually all supernatants were mixed together and 25 L aliquots were placed in autosampler at 4°C.

Untargeted LCMS Analysis

Analyses were performed using an Agilent 1290 UPLC coupled to a 6538 UHD Accurate-Mass Q-TOF (Agilent Technologies). The system was operated in positive electrospray ionization mode. Vials containing extracted metabolites and standard mixtures were kept at -80°C prior to LCMS analysis. *S. solfataricus* metabolites were separated using a Phenomenex C18 column (Kinetix 1.7 μ m, 100Å, 150mm x 2.1mm) using a flow rate of 500l/min. *S. scrofa* metabolites were separated using an Agilent C18 column (Zorbax RRHD Eclipse Plus 1.8 μ m, 150mm x 2.1mm) using a flow rate of 700l/min. In positive ionization mode, A = 0.1% formic acid in water B = 0.1% formic acid in acetonitrile, temperature was maintained at 50 °C for separation. For *S. solfataricus*, a linear gradient was used 100% A (0 - 2 min) – 100%B with an injection volume of 15 L. For *S. scrofa* an optimized step gradient was used 2%B (0 – 2 min), 2 - 22%B (2 - 29min), 22 – 92%B (29 – 36 min), 92%B (36 – 37 min), with an injection volume of 20l. ESI conditions were modified from (Yanes et al., 2011) gas temperature 300°C, drying gas of 12 L/min, nebulizer 55 psig, fragmentor 120V, and skimmer 65V. The instrument was set to acquire over the m/z range 50-1300 with an acquisition of 1 spectra/second. Data was visualized and converted to mzdata format using MassHunter software package (Agilent Technologies). Data processing was done using XCMS software, small molecule characterizations were done using the Metlin database (Patti et al. 2012, Smith et al. 2005, Tautenhahn et al. 2008).

Principal Components Analysis and Hierarchical Clustering

Data visualization was done using principal component analysis (PCA) and hierarchical clustering (hc) to visualize the grouping of the samples. For each sample (*S. solfataricus* and *S. scrofa*) control and treatment, PCA analysis was done using R functions `prcomp` for calculation of principal components and `hc` with R function `hclust` for calculation of dendrogram (fig.1). PCA was visualized using `rgl` library (R Development Core Team et al. 2012).

Evaluation of Variation

To estimate the degree of biological variation or noise in samples of each datasets, the coefficient of variation (CV) of ion intensity was calculated for all identified metabolites across replicates in each experimental condition as the standard deviation divided by the mean. A frequency plot of the CV being calculated was created, one plot for each condition (fig. 2). This evaluation has three purposes. First, it reveals any imbalance in biological variation across experimental conditions being compared in a given dataset. Second, it reveals the degree of this biological variation. Third, it allows a direct comparison of the biological variation of distinct datasets, such as the *S. scrofa* hemorrhagic stress and *S. solfataricus* oxidative stress datasets.

SVM-RFE Data Processing

A reference implementation of the SVM-RFE algorithm, available from the Weka machine learning toolkit (Hall et al. 2010.) was used for this project (class `SVMAttributeEval`, used with default parameters). The underlying SVM classifier that

was used has a linear kernel as in (Guyon et al. 2002). In short, SVM-RFE is based on SVM, a set of related supervised learning methods used for classification and regression. Briefly, an SVM algorithm will represent objects as points from various classes in a n -dimensional space (in our case, objects are samples from either control or treatment condition, classes are either ‘control’ or ‘treatment’, and n is the number of metabolites identified in the corresponding dataset). SVM will then construct a hyperplane which maximizes the distance to the neighboring points of both classes. SVM-RFE works by iteratively training SVM to discriminate control from treatment samples, identifying the features (metabolites) most informative for this discrimination, removing features that are least informative, then repeating the process with the new, reduced dataset. The end result is an ordering of the metabolites from most to least discriminative of the two conditions. To assess the robustness of the SVM-RFE algorithm with respect to the number of replicates of control and/or treatment conditions, the following procedure was implemented. Starting with a set of 10 control samples and 10 treatment samples, the SVM-RFE algorithm was used to produce a ranked list of metabolites for all combinations of numbers of control and treatment samples from ten of each (20 samples total) down to one of each (2 samples total). Then, for any given number k of samples in either control or treatment conditions, all k -combinations of samples were explored. A total of 1,046,529 runs of the SVM-RFE, were performed on a high-performance 11-nodes computer cluster. Merit scores of all metabolites across all sets of control and condition samples were aggregated to report the mean and standard deviation (for any set (n, m) of numbers of control and sample samples, the mean merit and standard deviation

for any given metabolite was calculated by aggregating merits across all n -combinations of all m -combinations).

SVM-RFE Model Validation

In order to validate the classification models and subsequent compounds ranking produced by the various SVM-RFE runs, the top 1%, 5%, 10% and 20% of the compounds ranked by SVM-RFE were used to train a SVM classifier to discriminate between control and stress treatments. The class functions.SMO from Weka was selected as the reference implementation and used with default parameters. For comparison, an additional SVM training was performed using the original list of compounds (100%). Leave-one-out cross-validation was performed to prevent overtraining.

Results and Discussion

Metabolites were extracted from biological replicates of the archaeal microbe *Sulfolobus solfataricus* ($n = 10$) stressed with 30 M H_2O_2 and compared to baseline after 30 minutes (Heinemann et al. 2014). In a second experiment *Sus scrofa* urine samples ($n = 8$) were collected 2 hours after hemorrhagic shock and compared to baseline. Molecular features were extracted using XCMS (Tautenhahn et al. 2008). To allow the comparison of multiple replicates and experimental conditions, only those features present in all of the biological samples were considered (*S. solfataricus* = 795, *S. scrofa* = 1,312). Principal component analysis (PCA) grouping showed a degree of conservation in phenotypic response when comparing the stressed to baseline conditions for both

experiments, however hierarchical clustering showed grouping for *S. solfataricus* was much better than *S. scrofa* where the phenotypic response was more ambiguous (fig. 1).

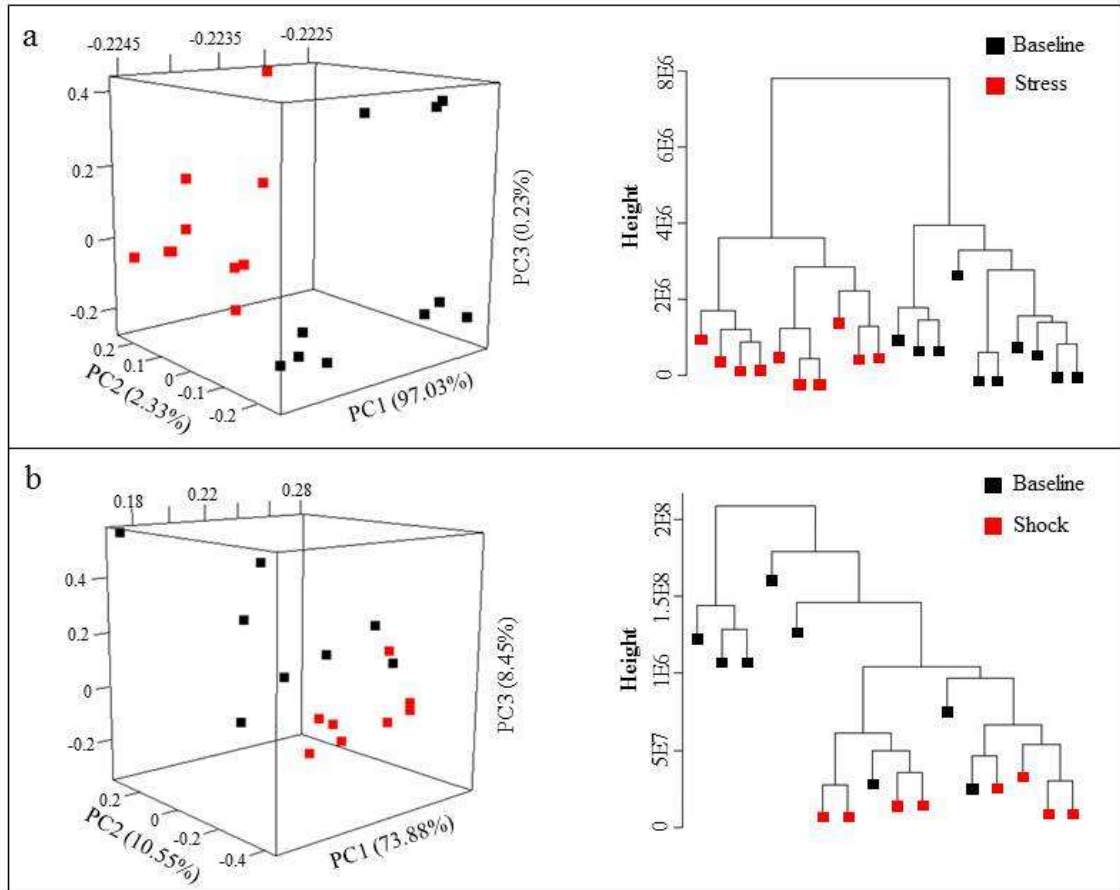


Figure 4.1 PCA Plots and Hierarchical Clustering. (a) ten biological replicates of *S. solfataricus* before and after 30 mM H_2O_2 stress and (b) eight biological replicates from *S. scrofa* before and after induction of hemorrhagic shock.

The coefficient of variation (CV) of relative abundance was calculated for each metabolite. The population CV for baseline and stress was plotted for each experimental condition (fig. 2). The average CV for the *S. solfataricus* population was less than half

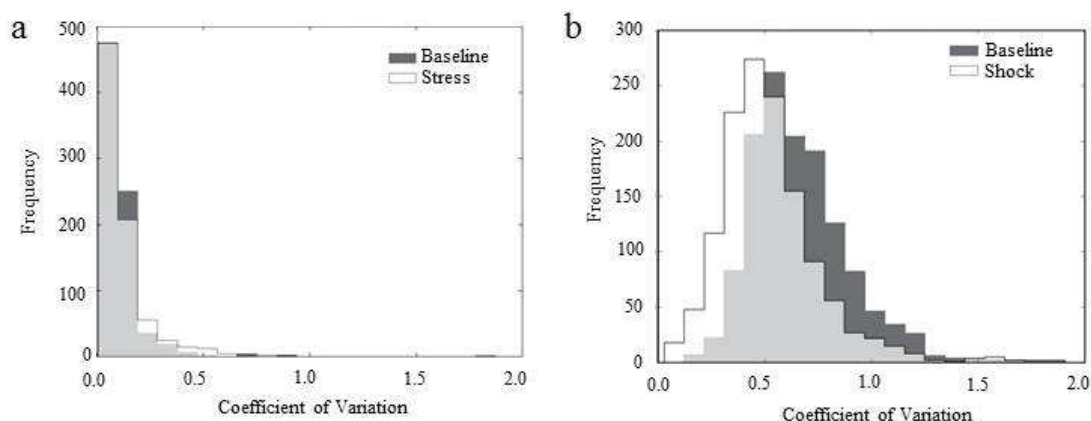


Figure 4.2 Coefficient of Variation (CV) Plots for Metabolites. (a) *S. solfataricus* and (b) *S. scrofa*, graphs show the CV in metabolite abundance. (a) *S. solfataricus* CV plots from baseline samples (gray) versus stressed (white), majority of features have low variance (CV < 0.1) (b) *S. scrofa* variation is considerably higher, (average ~ 0.5) shocked (white) have a distinct shift in CV from baseline (gray).

that of *S. scrofa*, indicating the two sample sets provide both a clean and noisy example of biological variation.

SVM-RFE was designed to output a rank for each metabolite based on the power to discriminate two sets of samples from distinct experimental conditions. A known problem in the application of machine learning is overtraining, which can occur when the number of features is large and the number of training patterns is comparatively small (Guyon et al. 2002). In contrast, SVM perform particularly well with a large number of features and a small number of training patterns (Guyon et al. 2002). However, relatively few examples are available for the analysis of untargeted metabolite data. Therefore, to evaluate how metabolite ranking was affected by the number of training replicates, relative intensity of metabolic features calculated using XCMS were directly imported into SVM-RFE via the Weka machine learning toolkit (Hall et al. 2010.). Starting with all available replicates we sequentially removed a replicate and evaluated how the

metabolite rank was affected (fig. 3). Using the SVM-RFE algorithm, a ranked list of metabolites together with merit scores was generated for all combinations of metabolite samples (i.e., sets of metabolite identifiers and corresponding abundances). This score quantified the ability of a given metabolite for the SVM algorithm to properly distinguish between the baseline and stressed conditions; the less the SVM algorithm relied on a metabolite to make this distinction, the lower its merit score. A common outcome was observed in both datasets (*S. solfataricus* and *S. scrofa*): the highest ranked metabolites maintained a consistent rank as the number of training replicates was reduced (fig. 3). This provides evidence top ranked features determined using few replicates can be robust predictors even in larger populations. However, biological variation had a clear impact on the ability of the SVM to consistently rank metabolites. In the *S. solfataricus* samples, metabolites retained their rank better than *S. scrofa* (fig. 3A, B). Due to less variation in *S. solfataricus* metabolite abundance, classification accuracy was better maintained when the number of training replicates was reduced (fig 3C, D). However, even in the *S. scrofa* experiment after SVM-RFE was used to reduce the features, classification accuracy could be maintained at 100% after cross-validation when $n \geq 3$ (Table 2).

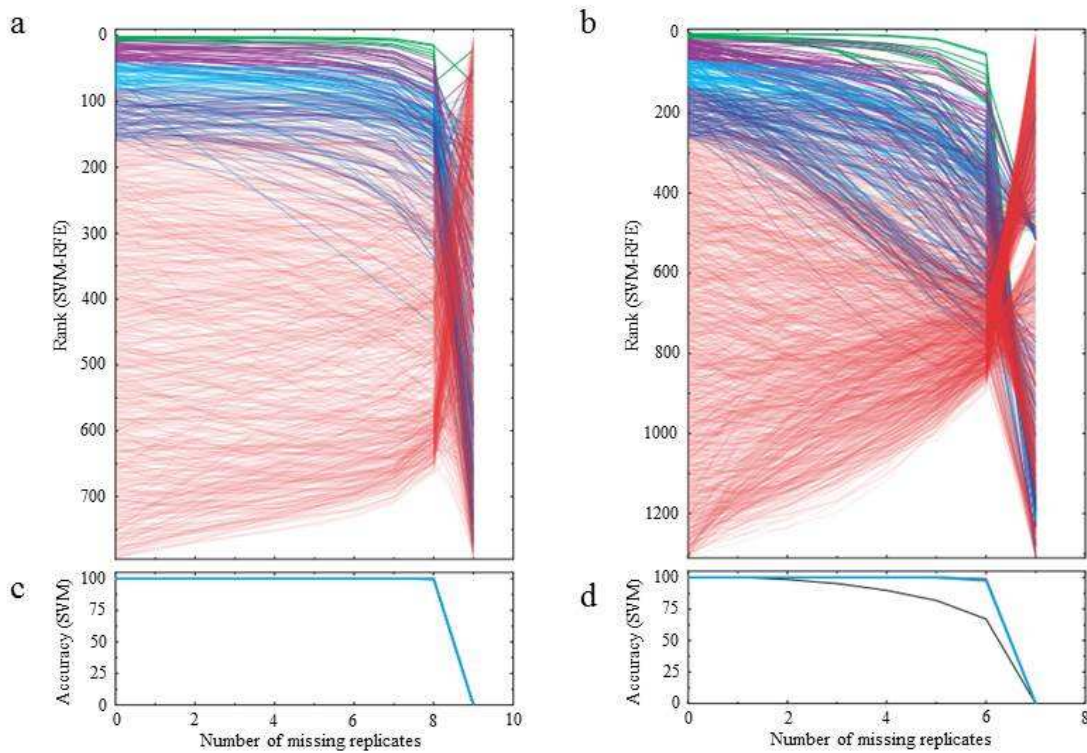


Fig. 4.3 Effect of Missing Replicates on Feature Rank and Accuracy. Each line represents an individual metabolite. The different colors were used to help visualize feature rank when reducing the number of training replicates (top ranked features, green:1%, violet: 5%, light blue: 10%, dark blue: 20%) (Y-axis) rank of metabolites (X-axis) number of replicates omitted. Above: (a) Rank of metabolites found in *S. solfataricus* (n=10), feature rank is consistent even after reducing the number of replicates (>2). (b) Rank of metabolites found in *S. scrofa* (n = 8), after reducing the number of replicates the rank is affected dramatically. SVM accuracy with reduced amounts of training replicates in *S. solfataricus* (c) and *S. scrofa* (d) grey line represents classification accuracy when all features are used. Accuracy appears the same in *S. solfataricus* whether all features or reduced numbers are used.

Table 4.1 Top ten features as determined by SVM-RFE

<i>S. solfataricus</i>				<i>S. scrofa</i>			
Rank	m/z	Fold Change	P-value	Rank	m/z	Fold Change	P-value
1	116.0694	4.84	0	1	192.0240	5.47	2.53E-7
2	70.0649	4.41	0	2	361.0597	6.84	3.29E-5
3	129.0427	4.82	0	3	763.3061	2.73	1.02E-5
4	86.0937	2.36	1.55E-12	4	303.1190	2.66	1.32E-3
5	210.0591	53.28	9.08E-12	5	546.0043	39.2	4.01E-4
6	118.0849	1.47	4.29E-10	6	328.1172	2.94	2.15E-3
7	157.0310	2.70	0	7	377.0165	8.51	2.33E-5
8	458.1829	4.16	3.15E-12	8	427.1115	4.15	1.79E-3
9	284.0964	2.40	0	9	589.2185	2.38	4.53E-3
10	423.0773	1.92	1.64E-12	10	246.1489	3.85	3.02E-4

Table 4.2 Training accuracy (%) of feature subsets when reducing the number of biological replicates

S. solfataricus

Features (%)	Number of Missing Replicates										
	0	1	2	3	4	5	6	7	8	9	
1	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0
5	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0
10	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.95	0
20	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	0
100	100.00	100.00	100.00	100.00	100.00	100.00	100.00	99.99	99.22		0

S. scrofa

Features (%)	Number of Missing Replicates							
	0	1	2	3	4	5	6	7
1	100.00	100.00	100.00	100.00	100.00	100.00	98.60	0
5	100.00	100.00	100.00	100.00	100.00	99.94	98.47	0
10	100.00	100.00	100.00	100.00	100.00	99.92	97.35	0
20	100.00	100.00	100.00	100.00	99.99	99.54	93.75	0
100	100.00	100.00	98.40	95.23	89.78	81.77	67.03	0

The urine samples provided a good example of biological worst case scenario as metabolite abundance in these samples is extremely noisy in comparison to other biological fluids (Veselkov et al. 2011). In both datasets, when using all features for classification, there is a reduction in accuracy if the number of training examples is reduced; however, if SVM-RFE is used to filter the number of features, classification accuracy can be maintained.

Concluding Remarks

Reliable identification of accurate biomarkers in metabolomics data can often be problematic as the number of biological replicates is often small and the degree of variation can be high. SVMs have been successful at selecting accurate biomarkers from noisy data in the past (Guyon et al. 2006). However, when using machine learning for biomarker selection in metabolomics studies, there is risk of overfitting because of the limited number of training examples and large number of metabolites. Here we show SVM-RFE has the ability to consistently identify accurate biomarkers even when trained with a limited number of biological samples (≥ 3). Our data shows that the top ranked biomarkers remain robust even when the training data has high variation. Further improvements are possible; indeed, SVM-RFE evaluates each biomarker independently and as such may miss functional relationships among compounds (e.g., metabolites involved in the same pathways). Other feature selection techniques such as SVM-based forward selection (starting with a randomly selected compound and adding to it as long as the accuracy of the resulting classification improves) or backward elimination

(compounds are randomly removed followed by iterative testing for changes in the accuracy of classification) combined with a proper heuristic (Hall et al., 2010) will rank sets of biomarkers, thus potentially exploit information about existing functional modules in the cell. Still, the culmination of these results suggests that use of SVM-RFE for biomarker selection is highly appropriate and can be effectively applied to biological experiments when there are a limited number of sample replicates. It has been shown that SVM performs similar to other machine learning methods for feature selection including genetic algorithms and random forest (Lin et al. 2011), therefore these methods may be similarly appropriate in feature selection with limited replicates.

Acknowledgements

This work was supported by National Science Foundation, MCB0646499 and MCB102248. Mass Spectrometry, Proteomics and Metabolomics core facility supported by the Murdock Charitable Trust, INBRE MT Grant No. P20 RR-16455-08, NIH Grant Nos. P20 RR-020185 and P20 RR-024237 from the COBRE Program of the National Center for Research Resources.

References

- Bertini, I., Calabro, A., Carli, V. De, Luchinat, C., Nepi, S., Porfirio, B., Renzi, D., et al. (2009). The Metabonomic Signature of Celiac Disease. *Journal of Proteome Research*, 8, 170–177.
- Duan, K., Rajapakse J.C., et al. (2005). Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data. *IEEE Transactions on Nanobioscience*, 4, 228-234.
- Guan, W., Zhou, M., Hampton, C. Y., Benigno, B. B., Walker, L. D., Gray, A., McDonald, J. F., et al. (2009) Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics*, 10, 259.
- Guyon I, Weston J, Barnhill S, (2002) Gene Selection for cancer classification using support vector machines, *Machine learning*. 46, 389-422.
- Hall M, National H, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH, et al. (2013) *The WEKA Data Mining Software* □: An Update, 11(1), 10–18.
- Heinemann J, Hamerly T, Maaty WS, Movahed N, Steffens JD, Reeves BD, Hilmer, JK, Therien J, Grieco PA, Peters JW, Bothner B, (2014) Expanding the paradigm of thiol redox in the thermophilic root of life. *Biochimica et Biophysica Acta* 1840, 80-85.
- Herder, C., Karakas, M., & Koenig, W., et al. (2011). Biomarkers for the prediction of type 2 diabetes and cardiovascular disease. *Clinical Pharmacology and Therapeutics*, 90(1), 52–66.
- Lin X, Wang Q, Yin P, Tang L, Tan Y, Li H, Yan K, (2011) A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. *Metabolomics*, 7(4), 549–558.
- Luszczek E R, Nelson T, Lexcen D, Witowski NE, Mulier KE, (2011) Urine Metabolomics in Hemorrhagic Shock: Normalization of Urine in the Face of Changing Intravascular Fluid Volume and Perturbations in Metabolism. *Bioanalysis & Biomedicine*, 3(2) 38-48.
- Maaty WS, Bothner B, Something old, something new, something borrowed; how the thermoacidophilic archaeon *Sulfolobus solfataricus* responds to oxidative stress, *PloS one*. 4 (2009) e6964.

Mahadevan S, Shah SL, Marrie TJ, Slupsky CM, (2008). Analysis of metabolomic data using support vector machines. *Analytical Chemistry*, 80(19), 7562–70.

Mulier KE, Beilman GJ, Conroy MJ, Taylor JH, Skarda DE, (2005). Ringer's ethyl pyruvate in hemorrhagic shock and resuscitation does not improve early hemodynamics or tissue energetics. *Shock*, 23, 248-252

Patti GJ, Tautenhahn R, Siuzdak G, (2012). Meta-analysis of untargeted metabolomic data from multiple profiling experiments. *Nature Protocols*, 7(3), 508–16.

Patti GJ, Yanes O, Shriver LP, Courade J, Tautenhahn R, Manchester M, Siuzdak G, (2012). Metabolomics implicates altered sphingolipids in chronic pain of neuropathic origin. *Nature Chemical Biology* 8(3), 232–234.

R Development Core Team (2012). R: A language and environment for statistical computing, reference index version 2.15.1. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>

Scribner DM, Witowski NE, Mulier KE, Luszczek ER, Wasiluk KR, Bielman GJ, (2010) Liver metabolomic changes identify biochemical pathways in hemorrhagic shock. *The Journal of Surgical Research*, 164, e131-9.

Serkova NJ, Standiford TJ, Stringer KA, (2011) The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses. *American Journal of Respiratory and Critical Care Medicine*. 184, 647–55.

Smith C, O'Maille G, Want EJ, Qin C, Siuzdak G, METLIN: a metabolite mass spectral database, Therapeutic drug monitoring. 27 (2005) 747–51.

Tautenhahn R, Böttcher C, Neumann S, (2008) Highly sensitive feature detection for high resolution LC/MS, *BMC Bioinformatics*. 9, 504.

Veselkov KA, Vingara LK, Masson P, Robinette SL, Want E, Li JV, Barton RH, Boursier-Neyret C, Walther B, Ebbels TM, Pelczer I, Holmes E, Lindon JC, Nicholson JK, (2011). Optimizing preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry*, 83, 5864-72

Yanes O, Tautenhahn GJ, Patti G, Siuzdak G, (2011). Expanding coverage of the metabolome for global metabolite profiling, *Analytical chemistry*. 83, 2152–61.

CHAPTER FIVE

REAL-TIME DIGITIZATION OF METABOLOMIC PATTERNS FROM A LIVING
SYSTEM USING MASS SPECTROMETRYContribution of Authors and Co-Authors

Manuscript in Chapter 5

Author: Joshua Heinemann

Contribution: Culturing of E.coli; Designed and fabricated microfluidics chip; Designed and set up robot; Formatted data; Designed experiments; Procured funding; Wrote manuscript.

Co-Author: Brigit Noon

Contribution: Made microfluidics chips; Assisted with system setup; Cultured E. coli; Ran many preliminary tests; Procured funding; Edited manuscript.

Co-Author: Mohammad J. Mohigmi

Contribution: Assisted with microfluidics chip design; Trained J. Heinemann for microfabrication.

Co-Author: Aurélien Mazurie

Contribution: Provided valuable insight; Wrote manuscript.

Co-Author: David L. Dickensheets

Contribution: Procured funding; Provided valuable insight.

Co-Author: Brian Bothner

Contribution: Provided valuable insight; Wrote manuscript; Procured funding.

Manuscript Information Page

Authors: Joshua Heinemann, Timothy Hamerly, Walid S. Maaty, Navid Movahed, Joseph D. Steffens, Benjamin D. Reeves, Jonathan K. Hilmer, Jesse Therien, Paul A. Grieco, John W. Peters and Brian Bothner

Journal: Journal of the American Chemical Society

Status of the manuscript:

- ☒ Prepared for submission to a peer-reviewed journal
☐ Officially submitted to a peer-reviewed journal
☐ Accepted by a peer-reviewed journal
☐ Published in a peer-reviewed journal

Publisher: American Chemical Society

REAL TIME DIGITIZATION OF METABOLOMIC PATTERNS FROM A LIVING SYSTEM USING MASS SPECTROMETRY

Joshua Heinemann¹, Brigit Noon¹, Mohammad J. Mohigmi², Aurélien Mazurie³, David L. Dickensheets² and Brian Bothner^{1,4}

Department of chemistry and biochemistry¹, Electrical & computer engineering department², Bioinformatics core facility³, Montana microfabrication facility⁴, Mass spectrometry, metabolomics and proteomics core facility⁵ Montana State University, Bozeman, MT 59717

Abstract

The real-time quantification of changes in intracellular metabolic activities has the potential to vastly improve upon traditional transcriptomics and metabolomics assays for the prediction of current and future cellular phenotypes. This is in part because intracellular processes reveal themselves as specific temporal patterns of variation in metabolite abundance that can be detected with existing signal processing algorithms. While metabolite abundance levels can be quantified by mass spectrometry (MS), large-scale real-time monitoring of metabolite abundance has yet to be realized due to technological limitations for fast extraction of metabolites from cells and biological fluids. To address this issue, we have designed a microfluidic-based inline small molecule extraction system, which allows for continuous metabolomic analysis of living systems using MS. The system requires minimal supervision, and has been successful at

real-time monitoring of bacteria and blood. Feature-based pattern analysis of *E. coli* growth and stress revealed cyclic patterns and forecastable metabolic trajectories. Using these trajectories, future phenotypes could be inferred as they exhibit predictable transitions in both growth and stress related changes. Herein we describe an interface for tracking metabolic changes directly from blood or cell suspension in real-time.

Introduction

Evidence suggests metabolite abundance exhibits phase like behavior when studied over time (Schmidt et al. 2011). This is best described in yeast, where glycolytic and other general metabolite oscillations are well documented (Richard et al. 2003). Metabolic oscillations have also been observed in bacteria (Levering et al. 2013), plants (Farre et al. 2012, Henson et al. 1990), insects (Schwemmler et al. 1980), mice (Dahlgren et al. 2005) and humans (Iotta et al. 2010, Bergensten et al. 2002, Obrig et al. 2000). Despite the potential value to systems biology and a deeper understanding of metabolic networks in general, little effort has been made to observe these oscillations using traditional metabolomics. Current analyses generally focus on fold change values at distinct time points, similar to transcriptomics (mRNA) and proteomics (protein) studies (Patti et al. 2012, Macounova et al. 2000). Metabolite abundance can change rapidly in response to even small perturbations, suggesting that important metabolic regulatory changes occur that current experimental and equipment set ups cannot capture due to low time resolution. Generally speaking, time course metabolic experiments are several orders of magnitude slower than the capability of a modern mass spectrometer (MS) to

collect data. This is because the rate limiting step for long term metabolite monitoring is the extraction of metabolites and their subsequent computational interpretation.

The integration of lab on a chip technology (LOC) with mass spectrometry (MS-LOC) has the potential to revolutionize biology by directly integrating living systems with real-time computational analysis. LOC functions have the ability to extract and selectively separate metabolite and protein biomarkers from biological fluids for MS abundance measurement (Schilling et al. 2002). This makes MS-LOC an ideal candidate for real-time computational analysis of living systems, and while MS does not share the computational speed of microprocessors (GHz), it has mHz detection rates making it compatible with the current timescale of parallelized LOC valving (Melin et al. 2007).

Liquid chromatography mass spectrometry (LCMS) based untargeted metabolomic profiles provide an excellent source of phenotypic information, as phenotype specific patterns emerge from the quantification of thousands of molecular features which change in response to time and stress (Tautenhahn et al. 2012). The majority of LCMS features are not easily identified; however, they still carry reproducible biological information which can be identified with pattern recognition algorithms. These algorithms allow us to predictively interpret the relationship between abundance patterns and phenotypes. Using a standardized system, metabolomic patterns could be used to monitor phenotypic transitions in real-time. However, one technological breakthrough which has inhibited metabolomics for real-time physiological analysis is an automated system to continuously separate the metabolites from a physiological fluid.

Here, we have engineered a system for extracting small molecules from biologically relevant fluids using a microfluidic based metabolite extraction chip (MEC) which is directly coupled to a mass spectrometer. This technology facilitates continuous automated biomolecule extraction, dramatically increasing the temporal resolution compared to standard metabolomics studies (Berk et al. 2011). Analyses of cell growth and perturbation studies demonstrate the potential for metabolic forecasting using the MEC (Box et al. 1976). This will vastly increasing the information recovery from metabolomics experiments through the detection of oscillating metabolic transient states. The overall setup is semi-supervised and can be controlled wirelessly providing a framework for cloud based computational analysis and real-time predictive metabolomics.

Materials and Methods

Metabolite Extraction Chip

Rapid prototyping of the MEC was achieved using the method described by Duffy et al. 2008 (Duffy et al. 1998). Polycarbonate molds for polydimethylsiloxane (PDMS) (Sylgard 184, Dow Corning) chips were fabricated in a mini milling machine (Grizzly Industrial Inc. Model G8689). The PDMS layer for the MEC was mixed at a ratio of 10:1, and was cured in a vacuum oven for 4 hours at 65°C. MEC were then sealed with oxygen plasma (MARCH etcher, Nordson) to a glass slide. PEEK Tubing with 1/32" outer diameter 0.005 mm ID (Upchurch) was inserted and sealed into a modified 18 Ga needle (Supplemental Figure 1) using PDMS as glue, to interface with the MEC. The

design of the chips was inspired by Bhagat et al. 2008, but optimized for diffusional separation of small molecules from cellular lysate (Bhagat et al. 2010, Bhagat et al. 2008). The fabricated MEC consisted of a five loop logarithmic spiral microchannel with two inlets and two outlets (Figure 1). The spiral had an outer diameter of 17.78 mm, and an inner diameter of 7.62 mm. Each port was 2.54 mm in diameter, the inlet and outlet channels were 250 μm wide, with the central microchannel 500 μm in width. All port and channels heights were 10-12 μm .

Metabolite Samples

E. coli strain BL21(DE3) was grown in Luria Bertani (LB) media (10 g tryptone, 5 g yeast extract, 10 g NaCl, pH=7 per Liter) in batch mode at 25 C with a beginning volume of 49 ml. Cultures were started by addition of 500 μl of *E. coli* that were precultured to a cell density 2.96×10^8 cells/ml (Volkmer et al. 2011). Ovine whole blood was collected in a sterile vial spiked with (2.6 g/ml) EDTA to prevent coagulation and injected directly onto the chip without any additional processing.

Mass Spectrometry

The oxidized glutathione (GSSG) diffusion experiment was performed on an ion trap (6340, Agilent Technologies). Analyses were performed using an isocratic flow (1100 HPLC, Agilent Technologies) of 50 $\mu\text{l}/\text{min}$ (50% A = 0.1% formic acid in water, 50% B = 0.1% formic acid in acetonitrile). The system was operated in positive electrospray ionization mode. In MS mode, the maximum accumulation time was 500 ms, scan range of 50-2200 m/z, and averages = 3. For multiple reaction monitoring

(MRM), masses were selected and isolated for MS(n) with a width of 4.0. For *E. coli*, a direct injection of ~30-592 cells (OD) generated a satisfactory signal in an injection volume of ~2 μ l. ESI-TOF analyses were performed using an isocratic flow (1290 UPLC, Agilent Technologies) of 50 μ l/min (50% A = 0.1% formic acid in water, 50% B = 0.1% formic acid in acetonitrile) coupled to a Q-TOF (6538, Agilent Technologies). The system was operated in positive electrospray ionization mode, using extended dynamic range (50-1700 m/z), 1 scan per second. For *E. coli*, direct injection of ~148-2960 cells in an injection volume of ~10 μ l generated satisfactory signal, without saturating the chip or MS.

Interfacing the Metabolite Extraction Chip and Mass Spectrometer

MS for metabolomics is typically setup in tandem with liquid chromatography (LC) for separation of the complex mixtures of small molecules. However, our system utilizes an isocratic LC flow through the MEC at all times with periodic direct injections of sample into the continuously flowing stream. The speed, volume and periodicity of each injection can be adjusted using the syringe pump (NE-4000, New Era Pump Systems Inc.) whose control is synced with the valve (MXP-9900, Rheodyne) using an in house program designed with labview (National Instruments). Small molecule diffusion was controlled by adjusting the passive flow rate through the chip where slower flow rates allowed greater diffusion across the channel. The syringe pump valve combination allowed for metabolomics data to be generated in a continuous fashion from culture and directly from blood. The data manifested as ionization pulses in the MS, where pulse

intensity could be controlled by the size of injections. Pulses had a periodicity of five minutes which was precisely controlled by the injection cycles of syringe pump valve combination. Testing showed that with the current tubing and value set up the period could be decreased to 3 minutes without significant degradation in performance.

Data Analysis

Data was extracted using pymzML, a Python library for high-throughput bioinformatics on mass spectrometry data (Bald et al. 2012). MS pulses were approximately two minutes in length and the intensity of each m/z was summed for the duration of the pulse and placed into a list. For *E. coli*, m/z were manually validated with Masshunter Qual (Agilent). The m/z intensities were imported into R where data visualization was done using principal component analysis (PCA) or plot, R function prcomp and plot. Data was mean centered (using function prcomp or scale) to negate the difference in pulse ionization abundance.

Results and Discussion

An engineered system for automated fast extraction of metabolites from cellular lysate that can operate in a continuous mode has great potential to revolutionize metabolic study by allowing direct analysis of a living system with high time resolution. To realize this goal, a MEC that consisted of a five loop logarithmic spiral microchannel with two inlets and two outlets was fabricated (Figure 1). During operation, an isocratic flow of 50 $\mu\text{l}/\text{min}$ (50% acetonitrile, 0.1% formic acid) was continuously run through the MEC and into the MS, with periodic injections of sample into channel 2. The MEC took

advantage of low Reynolds numbers in the microfluidic environment, which enabled our system to diffusively extract small molecules from cell lysate without turbulent mixing (Atencia et al. 2005). As the injected sample traversed the MEC, diffusion of solute

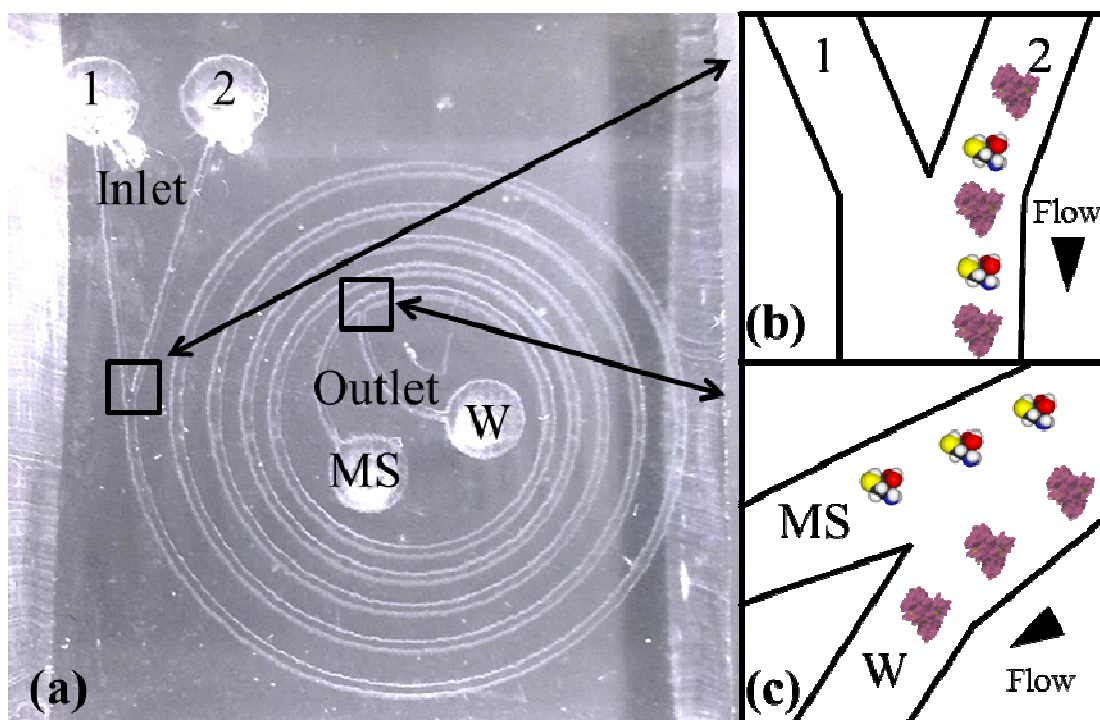


Figure 5.1 Diffusion of Small Molecules on Metabolite Extraction Chip (MEC). (a) PDMS chip, (b) inlet: *port 1*: 25% ACN, 0.1% FA; *port 2*: cell lysate and (c) outlet: MS: extracted metabolites to mass spectrometer and cellular debris to W: waste.

molecules from channel 2 to channel 1 occurred (Figure 1B). By design, only small molecules displayed rates of diffusion that allowed a significant accumulation in the channel directed for analysis. The outlet of the MEC sent the majority of the sample (cellular debris) to waste and the diffused small molecules directly to the MS for detection (Figure 1C). Channel diameter and flow was optimized to capture sufficient material for mass analysis, while minimizing buildup during long periods of continuous

instrument operation (>8 hrs). Injection of sample into the MEC was facilitated by a dual syringe pump which was placed in line with a 2-way 6 port valve (Figure 2). In valve position 1, syringe 1 was set

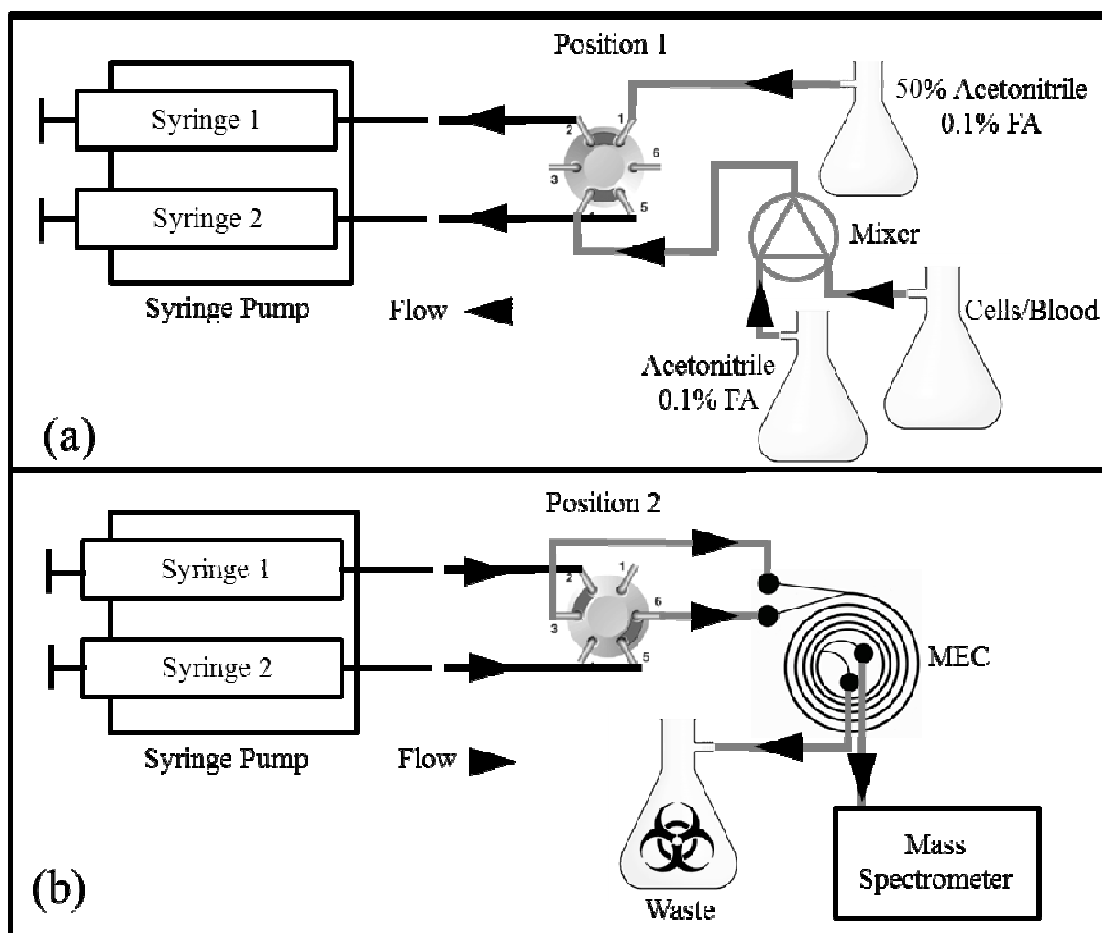


Figure 5.2 Binary Setup of Integrated Syringe Pump, Valve, Solvents, MEC, and Mass Spectrometer. (a) valve is in position 1, syringe pump is set to draw from cells and solvents; (b) valve is in position 2, syringe pump is set to pump

to draw MS solvent (50% acetonitrile, 0.1% formic acid) and syringe 2 the sample of interest (GSSG/E. coli/ovine blood) mixing with acetonitrile (Figure 2A). In position 2, the sample was injected into the respective channel of the MEC (Figure 2B). The

periodicity of injections resulted in pulses of metabolites which were easily detected by ESI-MS.

GSSG is a good molecular target for assessing cellular status as it is present in blood, during stress, disease and aging (Pastore et al. 2003). It is also a mid-sized metabolite with a molecular weight of 612 daltons (m/z 613.159 $M+H^+$) see Figure 3.

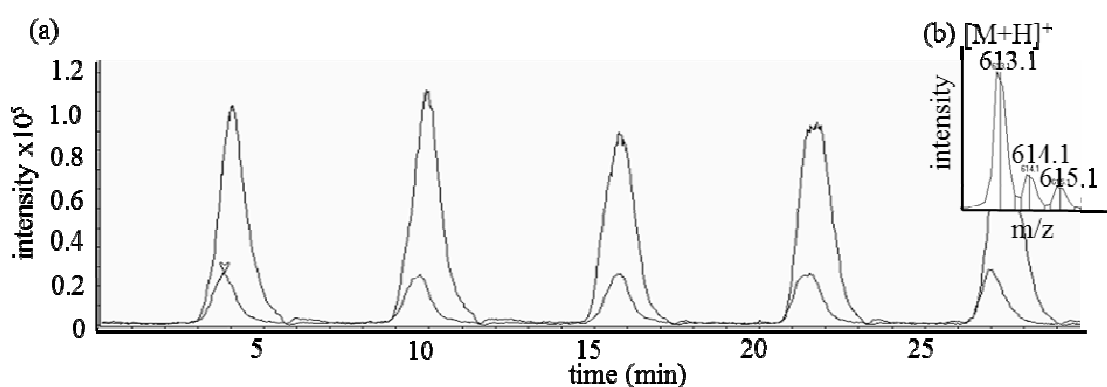


Figure 5.3 Diffusion of GSSG (613.1 m/z) in MEC on ESI-Trap. Injections are made every 5 minutes in a 30 minute run. (a) Extracted ion chromatogram (EIC) of GSSG. EIC of GSSG standard (higher) versus EIC of GSSG diffusion in MEC (lower). (b) Resolved isotopic peaks of GSSG.

Therefore, GSSG was used for testing performance of the MEC. Quantification of diffusion was measured by the abundance of GSSG detected in each of the outlet channels. These experiments used a set period of 5 min., and testing showed that the period could be decreased to 3 minutes without significant carryover. The MEC had a diffusion efficiency of 23.23 $\pm$ 2.33% for GSSG from channel 2 to channel 1 at 50 μ l/min in 50% acetonitrile with 0.1% formic acid. Testing showed this efficiency of small molecule diffusion to be the optimal, as it ensured that the vast majority of cellular polymers and lysis debris had insufficient time to diffuse and remained in the waste channel (Atencia et al. 2005) preventing buildup in the mass spectrometer.

To test the capabilities of our system for monitoring a living system, we tracked a growing culture of *E. coli*. Automated sampling directly from the cell culture on a five minute cycle was used. A special lysis chamber that used a voltage gap to lyse cells was designed and tested. After use it was noticed that the process of mixing the cell culture broth with acetonitrile resulted in efficient cell lysis (Supplemental Figure 2) so the additional module was removed. The cell lysate was then injected into the MEC for metabolite extraction and subsequent MS detection. Seventy-seven small molecules from media and cells could be detected consistently during the *E. coli* growth (Supplemental Figure 3). The relative abundance of select small molecules showed both smooth increases and decreases in abundance (Figure 4A) and oscillatory behavior during growth (Figure 4B). Because each of these small molecules displayed unique abundance changes over time, the oscillations were thought to be related to metabolic changes and not an instrument artifact. Interestingly the oscillations were not smooth, but bore a striking resemblance to the chaotic glycolytic oscillations observed in yeast. Using the abundance of the seventy-seven small molecules in each pulse, the changes over time were visualized using PCA (Figure 5). Time points were separated in chronological order, where each was most similar to those closest to it in time (Figure 5A, B). During one time course the system was paused and restarted after twenty minutes. This left a gap in the data which appears as a void in the PCA spacetime (Figure 5C). To provide evidence that the changes in metabolomic patterns were a product of growth, *E. coli* was allowed to progress to stationary phase and pulses were plotted by PCA (Figure 5D). The data now showed that sequential samples were no longer clearly distinguished by

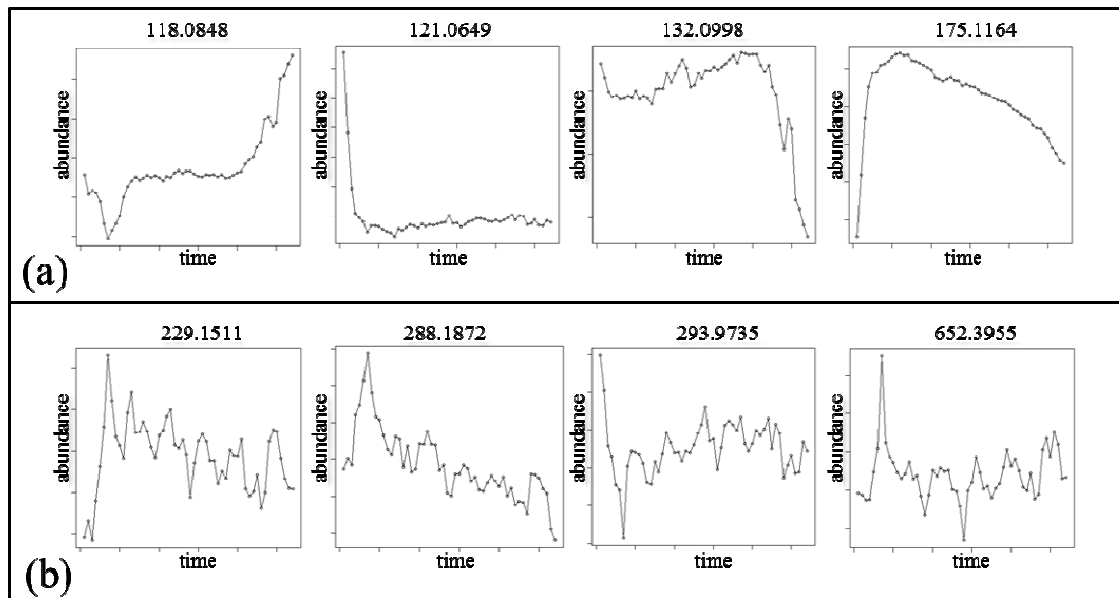


Figure 5.4 Change in Metabolite Abundance Over 6 hrs. of *E. coli* Growth, for Eight Small Molecules. (a) Abundance changes of four of small molecules which exhibit smooth transitions over time. (b) Abundance changes of four small molecules which exhibit oscillatory changes over time.

temporal linkage. Combined the PCA analysis revealed time dependent patterns could be clearly distinguished in the growing *E. coli* cultures.

Replicate cultures of the same cells under similar conditions would be expected to have similar metabolic fingerprints and somewhat reproducible changes over time. To test that this was in fact the case, data from the independent batch cultures of *E. coli* were analyzed together. A PCA plot of the data showed that there was distinct conservation in the trajectory of individual plots, however, the beginning and ending points differed (Figure 5E). This variation is probably due to the fact that cells used to inoculate the batch culture were taken from a continuous culture on different days. The goal of the experiments was not to minimize the variation between growths, but to instead observe the natural variation of the organism devoid of strict experimental control. In this way

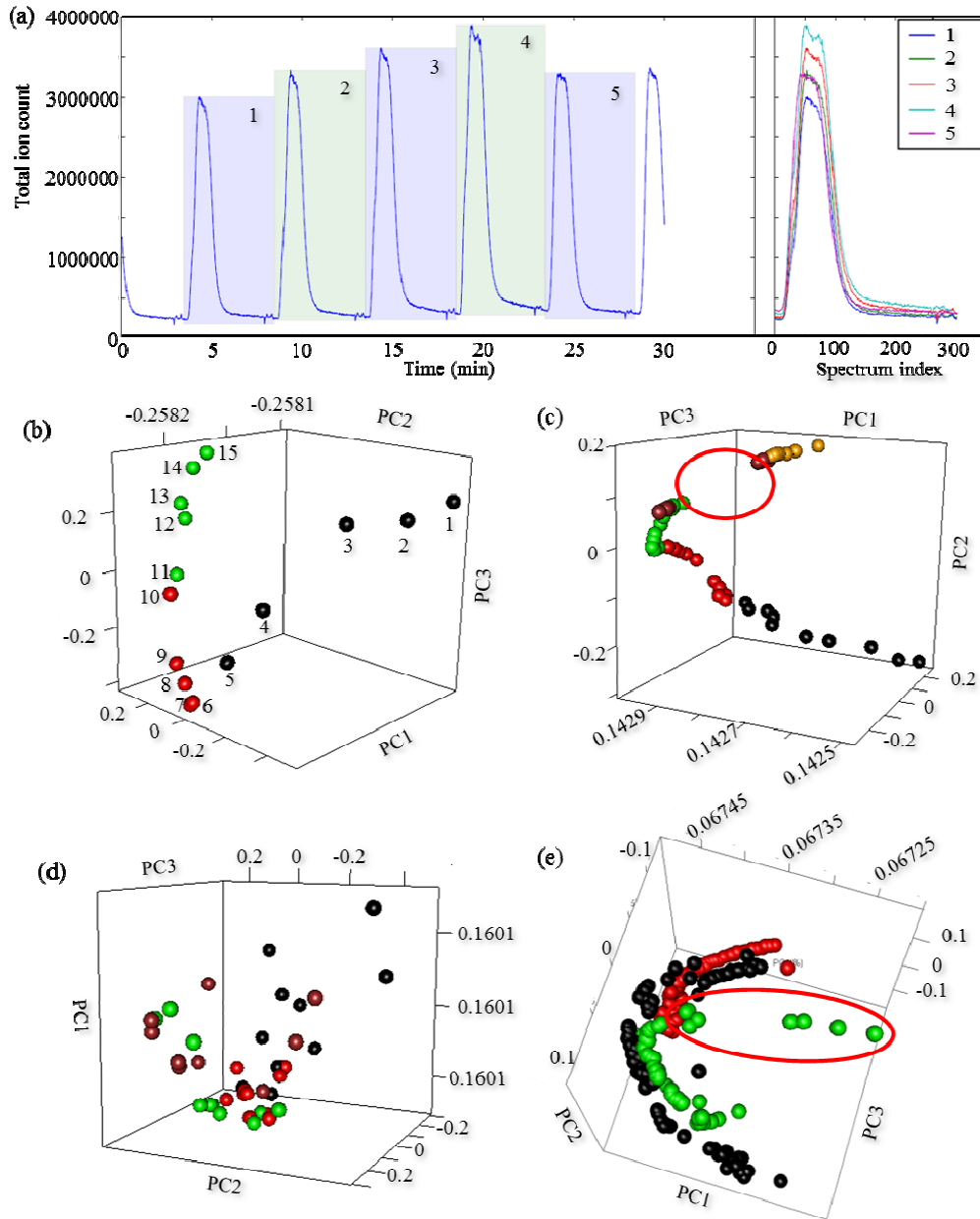


Figure 5.5 Real-Time Analysis of *E. coli* growth by MS. Colors of PCA indicate sequential subsets of time points. (a) Periodic injections of cellular lysate onto MEC yields total ion count (TIC) as the small molecules are detected by the MS (left). Difference in TIC depending on injection time point (right). (b) PCA of fifteen sequential time points during growth. (c) PCA of fifty time points during growth, red circle indicates missing time points when syringe pump ceased injections. (d) PCA of *E. coli* in stationary. (e) PCA of three growth replicates, each represented with a different color. (black) 8.5 hour growth, (red) 8.5 hour growth, (green) 6 hour growth, stressed with H_2O_2 after 4 hours, resulting in a different trajectory on the plot indicated with red circle.

the degree of conservation could be attributed to natural metabolic constraints not human experimental control. In order to test if small changes in the metabolic state of the cells could be detected, 100 μ M hydrogen peroxide was added to one replicate in the middle of the run. A change in the trajectory of the growth can be clearly seen (Figure 5E). The time points after the addition deviated from the smoothly curved overlaid trajectories before stress.

Plots in the abundance of the 77 small molecules over time were compared to find consistencies between the three replicates and difference which could be indicative of stress. Temporal changes in small molecule abundance were conserved across three different growths (Figure 6). This explains the similar trajectories presented in the PCA

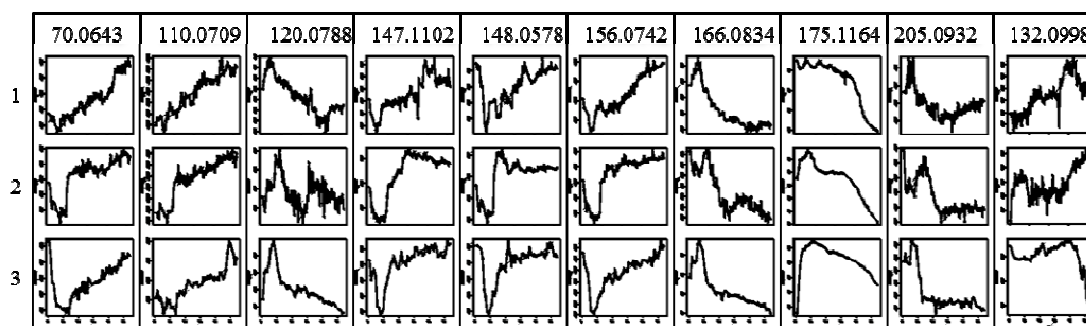


Figure 5.6 Change in Small Molecule Abundance During *E. coli* Growth for Ten Small Molecules in three growth replicates. Abundance changes for nine small molecules which exhibit similar changes over time with the exception of 132.0998 m/z where a decrease in abundance appears to occur in response to H_2O_2 stress.

above. However, markers of stress are also present. The feature at 132.0998 m/z was present in all growths, yet had a very different abundance after the addition of hydrogen peroxide. Combined, this evidence demonstrates this system is suitable for automated monitoring of metabolic processes. It is capable of quickly identifying growth patterns and phenotypic changes such as those caused by oxidative stress. Further development of

this technology could be especially useful for large scale monitoring of cell culture during mass production of pharmaceutical therapeutics and protein nanomaterials (Huang et al.2012).

The ability to monitor metabolic regulatory patterns could have implications far beyond the scope of tracking microbial growth phenotypes. This is because metabolomics abundance data can be highly sensitive to a wide variety of environmental influences including both disease and stress (Patti et al. 2012, Herder et al. 2011, Serkova et al. 2011). Therefore a miniaturized system such as the MEC used here could one day be used for extracting detailed information about the human body. Real-time monitoring of metabolic patterns from a population could allow preemptive identification of disease and/or monitoring of pharmaceutical intervention. To test the applicability of the current MEC setup and data analysis for mammalian metabolic monitoring, samples of whole blood were analyzed. The system injected 5 μ l samples of blood every five minutes and was successfully run for 1.5 hours without any problems (Supplementary Figure 4). Over one hundred small molecules could be detected. The system was capable of monitoring for extended periods of time, but because we were analyzing a blood sample that was previously collected, compared to an interavenous draw, no changes in metabolic activity were observed.

Concluding Remarks

Metabolomics technology is often used for identification of biomarkers that can be indicative of disease or stress. However, because the low time resolution of typical metabolomics data it is easy to see how transient changes in metabolism could cause erroneous biomarker identification. Here we have designed a system to vastly increase the time resolution of metabolomic study²³ by automating metabolite extraction from living cells. The results indicate we can identify oscillatory regulatory patterns in real-time (while the organism is growing) from as few as thirty cells. However, the system does have pitfalls. The direct injection of small molecules decreases the number of molecular features we can see. Typically hundreds to thousands of small molecules can be viewed with MS when chromatography is used, but our system is limited as we only detected between ~77-100 small molecules per injection. This shortcoming could be overcome by integrating fast chromatography between injections, as injections only take place approximately every 5 minutes. In addition, the system is very large would need to be miniaturized for portable use. Portable capability could be close however, as most functions already exist in miniaturized forms including microfluidic pumps, valves and mixers, and MEMS mass spectrometers (Syms et al. 2009). Here we have provided the first system for observing real-time metabolomic patterns from a living system and raise the bar for future metabolomics time course study.

Acknowledgements

This work was supported by National Science Foundation, MCB0646499, MCB102248, Kopriva Graduate Fellowship, and Howard Hughes Medical Institute (HHMI). Montana microfabrication facility (MMF). Mass Spectrometry, Proteomics and Metabolomics core facility supported by the Murdock Charitable Trust, INBRE MT Grant No. P20 RR-16455-08, NIH Grant Nos. P20 RR-020185 and P20 RR-024237 from the COBRE Program of the National Center for Research Resources.

References

- Atencia J, Beebe DJ, (2005) Controlled microfluidic interfaces, *Nature*. 437, 648-55.
- Bald T, Barth J, Niehues A, Specht M, Hippler M, Fufezan C, (2012) pymzml – Python module for high throughput bioinformatics on mass spectrometry data, *Bioinformatics*. 2012, 28(7), 1052-3.
- Bergsten P, Westerlund J, Liss P, Carlsson PO, (2002) Primary in vivo oscillations of metabolism in the pancreas, *Diabetes*. 51(3), 699-703.
- Berk M, Ebbels T, Montana G, (2011) A statistical framework for biomarker discovery in metabolomics time course data, *Bioinformatics*. 27(14), 1979-85.
- Bhagat AS, Kuntaegowdanahalli SS, Papautsky I, (2008) Continuous particle separation in spiral microchannels using dean flows and differential migration, *Lab Chip*. 8, 1906-14.
- Bhagat AS, Bow H, Wei Hou H, Jin Tan S, Han J, Lim C, (2010) Microfluidics for cell separation, *Med. Biol. Eng. Comput.* 48, 999-1014.
- Box GP, Jenkins GM, *Time series analysis forecasting and control*. Holden-Day Inc., San Francisco, 1976
- Dahlgren GM, Kauri LM, Kennedy RT, (2005) Substrate effects on oscillations in metabolism, calcium and secretion in single mouse islets of Langerhans, *Biochim. Biophys. Acta*. 1724, 23-26.
- Duffy DC, McDonald JC, Schueller JA, Whitesides GM, (1998) Rapid prototyping of microfluidic systems in poly(dimethylsiloxane), *Anal. Chem*. 70, 4974-84.
- Farre EM, Weise SE, (2012) The interactions between the circadian clock and primary metabolism, *Curr. Opin. Plant Biol*. 15(3), 293-300.
- Henson CA, Duke SH, (1990) Oscillations in plant metabolism, *Prog. Clin. Biol. Res*. 341B, 821-34.
- Herder C, Karakas M, Koenig W, Biomarkers for the prediction of type 2 diabetes and cardiovascular disease, *Clin. Pharmacol. Ther.* 2011, 90(1), 52-66.
- Huang C, Lin H, Yang XJ, (2012) Industrial production of recombinant therapeutics *Ind. Microbiol. Biotechnol*. 2012, 39, 383-399.

- Iotti S, Borsari M, Bendahan D, (2010) Oscillations in energy metabolism, *Biochim. Biophys. Acta.* 1797, 1353-61.
- Macounova K, Cabrera CR, Holl MR, Yager P, (2000) Generation of natural pH gradients in microfluidic channels for use in isoelectric focusing, *Anal. Chem.* 72, 3745-51.
- Melin J, Quake SR, (2007) Microfluidic large-scale integration: the evolution of design rules for biological integration, *Annu. Rev. Biophys. Biomol. Struct.* 2007, 36, 213-31.
- Obrig H, Neufang M, Wenzel R, Kohl M, Steinbrink J, Einhaupl K, Villringer A, (2000) Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults, *Neuroimage.* 12(6), 623-39.
- Patti GJ, Yanes O, Shriver LP, Courade J, Tautenhahn R, Manchester M, Siuzdak G, Metabolomics implicates altered sphingolipids in chronic pain of neuropathic origin, *Nat. Chem. Biol.* 2012, 8(3), 232-234.
- Pastore A, Federici G, Bertini E, Piemonte F, (2003) Analysis of glutathione: implication in redox and detoxification, *Clin. Chim. Acta.* 333, 19–39.
- Patti GJ, Tautenhahn R, Siuzdak G, (2012) Meta-analysis of untargeted metabolomic data from multiple profiling experiments, *Anal. Chem.* 7, 508-16.
- Richard P, (2003) The rhythm of yeast, *FEMS Microbiol. Rev.* 27, 547-57.
- Levering J, Kummer U, Becker K, Sahle S, (2013) Glycolytic oscillations in a model of a lactic acid bacterium metabolism, *Biophys. Chem.* 172, 53-60.
- Serkova NJ, Standiford TJ, Stringer KA, The emerging field of quantitative blood metabolomics of biomarker discovery in critical illnesses, *Am J Respir Crit Care Med.* 2011, 184, 647-55.
- Schilling EA, Kamholz AE, Yager P, (2002) Cell lysis and protein extraction in a microfluidic device with detection by a fluorogenic enzyme assay, *Anal. chem.* 74, 1798-804.
- Schmidt MD, Vallabhajosyula RR, Jenkins JW, Hood JE, Soni AS, Wikswo JP, Lipson H, (2011) Automated refinement and inference of analytical models for metabolic networks, *Phys. Biol.*, 8, 055011.
- Schwemmler W, Herrman M, (1980) Oscillation in the energy metabolism of the insect host symbiont. II Analysis of possible endogenous rhythms in both systems, *Cytobios.* 27, 193-208.

Syms RRA, Advances in microfabricated mass spectrometers, *Anal. Bioanal. Chem.* 2009, 393, 427-29.

Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G,(2012) XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data, *Anal. Chem.* 84, 5035-39.

Volkmer B, Heinemann M, (2011) Condition-dependent cell volume and concentration of *Escherichia coli* to facilitate data conversion for systems biology modeling, *PLoS One.* 6(7) e23126.

CHAPTER 6

CONCLUDING REMARKS

The results of this dissertation have reinforced the wealth of biological information which can be gained through the integration of multi-omics datasets. Here we describe the integration of these datasets with the model thermophile *S. solfataricus* which shows both convergent and divergent properties to organisms located in other domains of life. Interestingly, large parts of an organism's genome are thought to be the result of proviral integration and in *S. solfataricus* we see a portion of the genome which still remains similar to the virus. The provirus contains similarity to viruses from all domains of life, indicating viruses share a common gene pool. The sequence level similarity between the provirus major capsid protein and that of the encapsulin families seems to provide evidence of a viral origin for this family of proteins. Providing support to the theory, that many viral proteins have been adopted by the host for novel functions.

In addition, the study of oxidative stress of *S. solfataricus*, combining the proteomics, metabolomics and genomics of this organism, gives an example of a biomolecular system which has been adapted to serve a completely divergent role in thermophilic organisms. That is, typical thiol disulfide ratios are maintained to favor the reduced thiol which protects against oxygen free radicals. In thermophilic organisms such as *S. solfataricus* it appears the thiol/disulfide ratio is skewed to favor disulfides, a characteristic which appears to increase the stability of proteins at high temperatures.

The successful integration of omics technology has led to questions on how emerging omics technology may further revolutionize our insight of biological processes and thus improve healthcare and quality of life. Traditional omics experiments can be referred to as “fishing expeditions” where the induction of a cellular phenotype is caused by a perturbing agent, i.e. the oxidative stress of *S. solfataricus* or hemorrhagic shock in pigs. When analyzing the data we seek to find the regulated metabolite, proteins or transcripts, in order to find which biomolecules are up or down regulated in response to the stress. This strategy is useful for generating rudimentary insight into biological processes, however does not generate data which can quantify the relationships between interacting biomolecules. The development of real-time capability in computational systems begs the questions “can we observe metabolomics in real-time, and what would be the potential benefit of real-time computational analysis?” In this dissertation, I answer these questions. In the “*Application of support vector machines to metabolomics experiments with limited replicates*” I explore the information yield in metabolomics patterns. Here we see that even in inherently noisy untargeted profiles we can see conservation of metabolic response to stress. The important part is that we only need a few metabolic replicates to find these conserved features. This may not seem significant in its singularity however when you combine it with the results from the other publication, “*Real-time digitization of metabolomics patterns from a living system*” we have the ability to monitor patterns associated with disease or stress in real-time. Why is this significant, and what is the next step? Here is an example of a timing device for digitization of biological components. In the MEC system, the chip separates pulses of

microbial lysate or whole blood into binary pulses. The frequency of these pulses can serve as a uniform timing system for functional manipulation of biomolecules in a microfluidics chip, where each pulse is timed and can be processed with different fluidic functions following its transformation into a pulse. The frequency of these pulses can be modulated to ensure compatibility with a wide variety of biosensors. Therefore, the future of this research lies in the development of specific biosensors and the structuring of a kernel for assembling the data into a computational form.

REFERENCES CITED

- Albers SV, Birkeland N, Driessen AJ, Gertig S, Zaparty M, SulfoSYS (Sulfolobus Systems Biology): towards a silicon cell model for the central carbohydrate metabolism of the archaeon *Sulfolobus solfataricus* under temperature variation, *Biochem. Soc. Trans.* 37 (2009) 58–64.
- Allocati N, Federici L, Masulli M, Di Ilio C, Distribution of glutathione transferases in Gram-positive bacteria and Archaea, *Biochimie.* 94 (2012) 588–596.
- Augenlicht A, Wahrman MZ, Halsey H, Anderson L, Taylor J, Lipkin M, Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro, *Cancer Research.* 47 (1987), 6017–21.
- Barabasi AL, Albert R, (1999) Emergence of scaling in random networks, *Science.* 286, 509–512
- Barabasi AL, Oltvai ZN (2004) Network biology: understanding the cell's functional organization, *Nature reviews genetics.* 5, 101–113.
- Barik D, Baumann W T, Paul M R, Novak B, Tyson JJ (2010) A model of yeast cell-cycle regulation based on multisite phosphorylation, *Molecular systems biology.* 6, 405.
- Bar-Yam Y, (2004) Multiscale variety in complex systems. *Complexity.* 9, 37–45.
- Beeby M, O'Connor BD, Ryttersgaard C, Boutz D, Yeates TO, The genomics of disulfide bonding and protein stabilization in thermophiles, *PLoS Biol.* 3 (2005) e309.
- Beratan DN, Betts JN, and Onuchic JN, (1991) Protein electron transfer rates set by the bridging secondary and tertiary structure. *Science* 252, 1285-88.
- Bergesten P, (2002) Role of oscillations in membrane potential, cytoplasmic Ca²⁺, and metabolism for plasma insulin oscillations, *Diabetes.* 51, Suppl 1: S171-6.
- Berk M, Ebbels T, Montana G, (2011) A statistical framework for biomarker discovery in metabolomics time course data, *Bioinformatics.* 27(14), 1979-85.
- Bertini, I., Calabro, A., Carli, V. De, Luchinat, C., Nepi, S., Porfirio, B., Renzi, D., et al. (2009). The Metabonomic Signature of Celiac Disease. *Journal of Proteome Research*, 8, 170–177.
- Bhagat AS, Bow H, Wei Hou H, Jin Tan S, Han J, Lim C (2010) Microfluidics for cell separation, *Medical & biological engineering and computing.* 48, 999-1014.
- Canelas AB, Pierick A, Ras C, Siefar RM, Heijnen JJ, Quantitative evaluation of intracellular metabolite extraction techniques for yeast metabolomics, *Anal. Chem.* 81

(2009) 7379–89.

Chakravarthi S, Jessop CE, Bulleid NJ, (2006) The role of glutathione in disulphide bond formation and endoplasmic-reticulum-generated oxidative stress, *EMBO Rep.* 7 (2006) 271–275.

Chen K C, Calzone L, Csikasz-Nagy A, Cross F R, Novak B and Tyson J J (2004) Integrative analysis of cell cycle control in budding yeast, *Molecular biology of the Cell.* 15, 3841–62

Chu Y, Corey DR (2012). RNA sequencing: platform selection, experimental design, and data interpretation, *Nucleic acid theurapeutics.* 22, 271-4.

Cortes C, Vapnick V, (1995) Support-vector networks, *Machine learning.* 20, 273-97.

Cumming RC, Andon NL, Haynes PA, Minkyu P, Schubert D, Protein disulfide bond formation in the cytoplasm during oxidative stress, *J. Biol. Chem.* 279 (2004) 21749–21758

Dahlgren GM, Kauri LM, Kennedy RT, (2005) Substrate effects on oscillations in metabolism, calcium, and secretion in single mouse islets of Langerhans, *Biochimica et biophysica acta.* 1724(1-2), 23-26.

Davidson EH, Rast JP, Oliveri P, Ransick A, Calestani C, Yuh CH, Minokawa T, Amore G, Hinman V, Arenas-Mena C, Otim O, Brown CT, Livi CB, Lee PY, Revilla R, Rust AG, Pan ZJ, Schilstra MJ, Clarke PJC, Arnone MI, Rowen L, Cameron RA, McClay DR, Hood L, Bolouri H (2002) A genomic regulatory network for development, *Science.* 295, 1669–78.

Derman J, Beckwith J, (1991) Escherichia coli alkaline phosphatase fails to acquire disulfide bonds when retained in the cytoplasm, *J. Bacteriol.* 173, 7719–7722.

Dlakić M, HHsvm: fast and accurate classification of profile-profile matches identified by HHsearch, *Bioinformatics.* 25 (2009) 3071–3076.

Duan, K., Rajapakse J.C., et al. (2005). Multiple SVM-RFE for Gene Selection in Cancer Classification With Expression Data. *IEEE Transactions on Nanobioscience*, 4, 228-234.

Fahey RC, (2001) Novel thiols of prokaryotes, *Annu. Rev. Microbiol.* 55, 333–356.

Farre EM, Weise SE, (2012) The interactions between the circadian clock and primary metabolism, *Current opinion in plant biology.* 15(3), 293-300.

Feng X, Xu Y, Chen Y, Tang YJ, (2012) Integrating flux balance analysis in kinetic

models to decipher the dynamic metabolism of *Shewanella oneidensis* MR-1, *PLoS Computational Biology*. 8, e1002376.

Fiehn O (2002). Metabolomics - the link between genotypes and phenotypes. *Plant molecular physiology*. 48 155-171.

Fratelli M, (2005) Gene expression profiling reveals a signaling role of glutathione in redox regulation, *Proc. Natl. Acad. Sci. USA*. 102, 13998–14003.

Gao D, Liu H, Jiang Y, Lin J (2013) Recent advances in microfluidics combined with mass spectrometry: technologies and applications, *Lab on a chip*. 13, 3309-22.

Gedeon T, Bokes P (2012). Delayed protein synthesis reduces the correlation between mRNA and protein fluctuations, *Biophysical journal*. 3, 377-85.

Grogan, D. W. (1989) Phenotypic Characterization of the Archaeobacterial Genus *Sulfolobus* - Comparison of 5 Wild-Type Strains, *Journal of Bacteriology*. 171, 6710-6719.

Guan, W., Zhou, M., Hampton, C. Y., Benigno, B. B., Walker, L. D., Gray, A., McDonald, J. F., et al. (2009) Ovarian cancer detection from metabolomic liquid chromatography/mass spectrometry data by support vector machines. *BMC Bioinformatics*, 10, 259.

Guyon I, Weston J, Barnhill S, (2002) Gene Selection for cancer classification using support vector machines, *Machine learning*. 46, 389-422.

Hall M, National H, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH, et al. (2013) *The WEKA Data Mining Software*: An Update, 11(1), 10–18.

Hayden EC, (2014) Is the \$1000 genome for real? *Nature*. doi:10.1038

Hansen RE, Roth D, Winther JR, Quantifying the global cellular thiol-disulfide status, *Proc. Natl. Acad. Sci. USA*. 106 (2009) 422–427.

Heinemann J, Maaty WS, Gauss GH, Akkaladevi N, Bothner B, Fossil record of an archaeal HK97-like provirus, *Virology*. 417 (2011) 362–8. *biological research*, 341B, 821-34.

Heinemann J, Hamerly T, Maaty WS, Movahed N, Steffens JD, Reeves BD, Hilmer, JK, Therien J, Grieco PA, Peters JW, Bothner B, (2014) Expanding the paradigm of thiol redox in the thermophilic root of life. *Biochimica et Biophysica Acta* 1840, 80-85.

Henson CA, Duke SH, (1990) Oscillations in plant metabolism, *Prog. Clin. Biol. Res.* 341B, 821-34.

Herder, C., Karakas, M., & Koenig, W., et al. (2011). Biomarkers for the prediction of type 2 diabetes and cardiovascular disease. *Clinical Pharmacology and Therapeutics*, 90(1), 52–66.

Huang, S., Wikswo, J., (2006) Dimensions of systems biology. *Reviews of physiology biochemistry and pharmacology* 157, 81-104.

Hummel CS, Lancaster KM, Edward J, Crane EJ, (2005) Determination of coenzyme A levels in *Pyrococcus furiosus* and other Archaea: implications for a general role for coenzyme A in thermophiles, *FEMS Microbiol. Lett.* 252, 229–234.

Iotti S, Borsari M, Bendahan D, (2010) Oscillations in energy metabolism, *Biochimica et biophysica acta.* 1797(8), 1353-61.

Jorda J, Yeates TO, (2011) Widespread disulfide bonding in proteins from thermophilic archaea, *Archaea.* 409156.

King RD, Rowland J, Oliver SG, Young M, Aubrey W, Byrne E, Liakata M, Markham M, Pir P, Soldatova LN, Sparkes A, Whelan KE, Clare A, (2009) The automation of science, *Science.* 324, 85-9.

Kitano, H., (2002) Systems biology: a brief overview. *Science* 295, 1662-1664.

Kuttykrishnan S, Sabina J, Langton LL, Johnston M and Brent MR, (2010) A quantitative model of glucose signaling in yeast reveals an incoherent feed forward loop leading to a specific, transient pulse of transcription, *Proceeding of the national academy of science of the United States of America.* 107, 16743–8.

Ladenstein L, Ren B, (2008) Reconsideration of an early dogma, saying "there is no evidence for disulfide bonds in proteins from archaea", *Extremophiles.* 12, 29-38.

Lamble HJ, Theodossis A, Milburn CC, Taylor GL, Danson MJ, (2005) Promiscuity in the part-phosphorylative Entner-Doudoroff pathway of the archaeon *Sulfolobus solfataricus*, *FEBS Lett.* 579, 6865–6869.

Larson ET, Eilers B, Menon S, Reiter D, Lawrence CM, (2007) A winged-helix protein from *Sulfolobus* turreted icosahedral virus points toward stabilizing disulfide bonds in the intracellular proteins of a hyperthermophilic virus, *Virology.* 368, 249–261.

Lawrence CM, Menon S, Eilers BJ, Bothner B, R. Khayat R, Douglas T, Young MJ, (2009) Structural and functional studies of archaeal viruses, *J. Biol. Chem.* 284, 12599-603.

Levering J, Kummer U, Becker K, Sahle S, (2013) Glycolytic oscillations in a model of lactic acid bacterium metabolism, *Biophysical Chemistry*. 172, 53-60.

Lin X, Wang Q, Yin P, Tang L, Tan Y, Li H, Yan K, (2011) A method for handling metabonomics data from liquid chromatography/mass spectrometry: combinational use of support vector machine recursive feature elimination, genetic algorithm and random forest for feature selection. *Metabolomics*, 7(4), 549–558.

Luszczek E R, Nelson T, Lexcen D, Witowski NE, Mulier KE, (2011) Urine Metabolomics in Hemorrhagic Shock: Normalization of Urine in the Face of Changing Intravascular Fluid Volume and Perturbations in Metabolism. *Bioanalysis & Biomedicine*, 3(2) 38-48.

Maaty WS, Heinemann J, Bothner B, (2009) Something old, something new, something borrowed; how the thermoacidophilic archaeon *Sulfolobus solfataricus* responds to oxidative stress, *PloS one*. 4, e6964.

Maaty WM, Selvig K, Ryder S, Tarlykov P, Hilmer JK, Heinemann J, Steffens J, Snyder JC, Ortmann AC, Movahed N, Spicka K, Chetia L, Grieco PA, Dratz EA, Douglas T, Young MJ, Bothner B. (2012) Proteomic analysis of *Sulfolobus solfataricus* during *sulfolobus* turreted icosahedral infection, *Journal of proteome research*. 11, 1420-32.

Maaty WS, Steffens J, Heinemann J, Ortmann AC, Reeves BD, Biswas SK, Dratz EA, Grieco PA, Young MJ, Bothner B, (2012) Global analysis of viral infection in an archaeal model system, *Front. Microbiol.* 3, 411.

Mallick P, Boutz DR, Eisenberg D, Yeates TO, Genomic evidence that the intracellular proteins of archaeal microbes contain disulfide bonds, *Proc. Natl. Acad. Sci. USA*. 99 (2002) 9679–9684.

Macounova K, Cabrera CR, Holl MR, Yager P, (2000) Generation of natural pH gradients in microfluidic channels for use in isoelectric focusing, *Analytical chemistry*. 72, 3745-51.

Mahadevan S, Shah SL, Marrie TJ, Slupsky CM, (2008) Analysis of metabolomic data using support vector machines. *Analytical Chemistry*, 80(19), 7562–70.

Malki L, Yanku M, Borovok Y, Cohen G, Aharonowitz Y, (2009) Identification and characterization of *gshA*, a gene encoding the glutamate-cysteine ligase in the halophilic archaeon *Haloferax volcanii*, *J. Bacteriol.* 191, 5196–5204.

- Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, la Favera R and Califano A (2006) ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context, *BMC Bioinformatics*. 7, Suppl. 1 S7.
- Maaty WS, Bothner B, Something old, something new, something borrowed; how the thermoacidophilic archaeon *Sulfolobus solfataricus* responds to oxidative stress, *PloS one*. 4 (2009) e6964.
- Melin J, and Quake SR, (2007) Microfluidic large-scale integration: the evolution of design rules for biological automation, *Annual review of biophysics and biomolecular structure*. 36, 213-31.
- Mulier KE, Beilman GJ, Conroy MJ, Taylor JH, Skarda DE, (2005). Ringer's ethyl pyruvate in hemorrhagic shock and resuscitation does not improve early hemodynamics or tissue energetics. *Shock*, 23, 248-252
- Menon SK, Maaty WS, Corn GJ, Bothner B, Lawrence CM, (2008) Cysteine usage in *Sulfolobus* spindle-shaped virus 1 and extension to hyperthermophilic viruses in general, *Virology*. 376, 270–278.
- Nakabachi A, Yamashita A, Toh H, Ishikawa H, Dunbar H, Moran N, Hattori M (2006) The 160-kilobase genome of the bacterial endosymbiont *Carsonella*, *Science*. 314, 267.
- Nunn CE, Johnsen U, Schonheit P, Fuhrer T, Danson MJ, Metabolism of pentose sugars in the hyperthermophilic archaea *Sulfolobus solfataricus* and *Sulfolobus acidocaldarius*, *J. Biol. Chem*. 285 (2010) 33701–33709.
- Obrig H, Neufang M, Wenzel R, Kohl M, Steinbrink J, Einhaupl K, Villringer A, (2000) Spontaneous low frequency oscillations of cerebral hemodynamics and metabolism in human adults, *Neuroimage*. 12(6), 623-39.
- Ohtomo, Y, Kakegawa, T, Ishida, A, Nagase, T, Rosing, MT, (2013) Evidence for biogenic graphite in early archaean isua metasedimentary rocks, *Nature Geoscience*. 7, 25-28.
- Orth JD, Thiele I, Palsson BO, (2010) What is flux balance analysis? *Nature Biotechnology* 28, 245-248.
- Ortmann, AC, Brumfield SK, Walther J, McInnerney K, Brouns SJ, Harmen, van de Werken, HJ, Bothner B, Douglas T, van de Oost J, Young MJ, Transcriptome analysis of the archaeon *Sulfolobus solfataricus* with *Sulfolobus* turreted icosahedral virus, *Journal of virology*. 82, 4874-83.

- Pai EF, Schulz GE, The catalytic mechanism of glutathione reductase as derived from x-ray diffraction analyses of reaction intermediates, *J. Biol. Chem.* 258 (1983) 1752-57.
- Pastore A, Federici G, Bertini E, Piemonte F, (2003) Analysis of glutathione: implication in redox and detoxification, *Clin. Chim. Acta.* 333, 19–39.
- Patti GJ, Tautenhahn R, Siuzdak G, (2012). Meta-analysis of untargeted metabolomic data from multiple profiling experiments. *Nature Protocols*, 7(3), 508–16.
- Patti GJ, Yanes O, Shriver LP, Courade J, Tautenhahn R, Manchester M, Siuzdak G, (2012). Metabolomics implicates altered sphingolipids in chronic pain of neuropathic origin. *Nature Chemical Biology* 8(3), 232–234.
- Patti GJ, Tautenhahn R, Rinehart D, Cho K, Shriver LP, Manchester M, Nikolskiy I, Johnson CH, Mahieu NG, Siuzdak G, (2013) A View from Above: Cloud Plots to Visualize Global Metabolomic Data, *Anal. Chem.* 85, 798-04.
- Patti GJ, Tautenhahn R, Siuzdak G, (2012) Meta-analysis of untargeted metabolomic data from multiple profiling experiments, *Nature Protocols*. 7508–16.
- Pellicer J, Fay MF, Leitch IJ, (2010) The largest eukaryotic genome of them all? *Botanical journal of the Linnean Society*. 164, 10-15.
- Richard P, (2003) The rhythm of yeast, *FEMS microbiology reviews*. 27, 547-57.
- R Development Core Team (2012). R: A language and environment for statistical computing, reference index version 2.15.1. *R Foundation for Statistical Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL <http://www.R-project.org>
- Rasband WS, ImageJ. U.S. National Institutes of Health, Bethesda, Maryland, USA. (2012) Available at: <http://rsbweb.nih.gov/ij/>.
- Richard P, (2003) The rhythm of yeast, *FEMS Microbiol. Rev.* 27, 547-57.
- Savageau MA, (2009) *Biochemical Systems Analysis: A Study of Function and Design in Molecular Biology*. Reprint of the Original Edition Published in 1976 (Reading, MA:Addison-Wesley).
- Schena M, Shalon D, Davis RW, Brown PO, (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science*. 270, 467-470.
- Schwemmler W, Herrman M, (1980) Oscillation in the energy metabolism of the insect host and symbiont II. Analysis of possible endogenous rhythms in both systems, *Cytobios*. 27, 193-208.

Schmidt MD, Vallabhajosyula RR, Jenkins JW, Hood JE, Soni AS, Wikswo JP, Lipson H, (2011) Automated refinement and inference of analytical models for metabolic networks, *Phys. Biol.*, 8, 055011.

Schwartzman D, Lineweaver CH, Biospheric Self-Organization A Hyperthermophilic Origin of Life, in: A. Kleidon, R.D. Lorenz, Non-equilibrium Thermodynamics and the production of Entropy. In Non-equilibrium Thermodynamics and the production of Entropy, Springer-Verlag Berlin Heidelberg, Germany, 2005, pp. 207-221.

Schwemmler W, Herrman M, (1980) Oscillation in the energy metabolism of the insect host symbiont. II Analysis of possible endogenous rhythms in both systems, *Cytobios.* 27, 193-208.

Scribner DM, Witowski NE, Mulier KE, Luszczek ER, Wasiluk KR, Bielman GJ, (2010) Liver metabolomic changes identify biochemical pathways in hemorrhagic shock. *The Journal of Surgical Research*, 164, e131-9.

Serkova NJ, Standiford TJ, Stringer KA, (2011) The emerging field of quantitative blood metabolomics for biomarker discovery in critical illnesses. *American Journal of Respiratory and Critical Care Medicine*. 184, 647–55.

Shah AR, Oehmen CS, Webb-Robertson B, (2008) SVM-HUSTLE--an iterative semi-supervised machine learning approach for pairwise protein remote homology detection, *Bioinformatics*. 24, 783–90.

She Q, Singh RK, Confalonieri F, Zivanovic, J. van der Oost J, (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2, *Proc. Natl. Acad. Sci. USA*. 98, 7835–7840.

Sevier CS, Kaiser C (2002) Formation and transfer of disulphide bonds in living cells, *National review molecular cell biology*. 3, 836–47.

Smirnova GV, Oktyabrsky ON, (2005) Glutathione in bacteria. *Biochemistry, Biokhimii* a. 70, 1199–1211.

Smith C, O'Maille G, Want EJ, Qin C, Siuzdak G, METLIN: a metabolite mass spectral database, *Therapeutic drug monitoring*. 27 (2005) 747–51.

Soldatova LN, Clare A, Sparkes A, King RD, (2006) An ontology for a robot scientist, *Bioinformatics*. 22, e464-71.

Soni AS, Jenkins JW and Sundaram S S (2008) Determination of critical network interactions: an augmented Boolean pseudo-dynamics approach, *IET Systems Biology*. 2, 55–63.

Steuer R, Junker BH, Computational models of metabolism: stability and regulation in metabolic networks, *Advances in chemical physics*. 142, ISBN-10: 0-470-46499-2.

Syms, RRA, (2009) Advances in microfabricated mass spectrometers, *Analytical and Bioanalytical Chemistry*. 393, 427-29.

Tautenhahn R, Böttcher C, Neumann S, (2008) Highly sensitive feature detection for high resolution LC/MS, *BMC Bioinformatics*. 9, 504.

Tautenhahn R, Patti GJ, Rinehart D, Siuzdak G, XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data, *Anal. Chem.* 84 (2012) 5035-39.

Tyson JJ, Chen KC, Novak B, (2003) Sniffers, buzzers, toggles and blinkers: dynamics of regulatory and signaling pathways in the cell *Current opinion in cell biology*. 15, 221–31.

Veselkov KA, Vingara LK, Masson P, Robinette SL, Want E, Li JV, Barton RH, Boursier-Neyret C, Walther B, Ebbels TM, Pelczar I, Holmes E, Lindon JC, Nicholson JK, (2011). Optimizing preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Analytical Chemistry*, 83, 5864-72

Voelkerding KV, Dames SA, Durtsche JD, (2009) Next-generation sequencing: from basic research to diagnostics, *Clinical chemistry*. 55, 641-58.

Voit EO, (1991) *Canonical Nonlinear Modeling: S-System Approach to Understanding Complexity* (New York: VanNostrand Reinhold)

Want EJ, O'maille G, Smith CA, Brandon TR, Uritboonthai W, Qin C, Trauger SA, Siuzdak G, (2006) Solvent-dependent metabolite distribution, clustering, and protein extraction for serum profiling with mass spectrometry, *Analytical chemistry*. 78, 743-52.

Weidenheft B, Mosolf J, Willits D, Yeager M, Dryden KA, Young M, Douglas T, (2005). An archaeal antioxidant: characterization of a Dps-like protein from *Sulfolobus solfataricus*, *Proceedings of the National Academy of Science of the United States of America*. 102(30), 10551-6.

Woese CR (2000). Interpreting the universal phylogenetic tree, *Proceedings of the National Academy of Sciences of the United States of America*. 97, 8392-96.

Yanes O, Tautenhahn GJ, Patti G, Siuzdak G, (2011). Expanding coverage of the metabolome for global metabolite profiling, *Analytical chemistry*. 83, 2152–61.

APPENDICES

APPENDIX A

SUPPORTING INFORMATION FOR CHAPTER 2

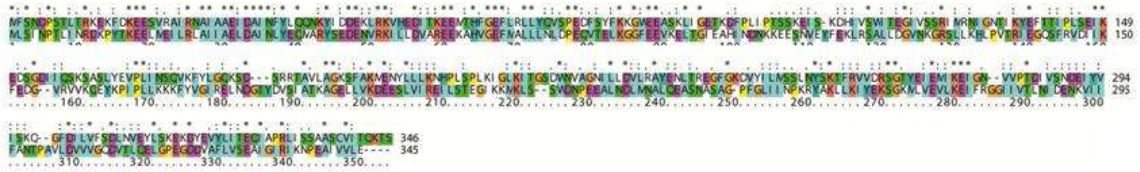


Figure S1: Alignment of hypothetical protein SSO2749 (GI:15899464) with Virus-like particle (GI:146387236) from *Pyrococcus furiosus*. * denotes identity; : and . denote homologous substitution.

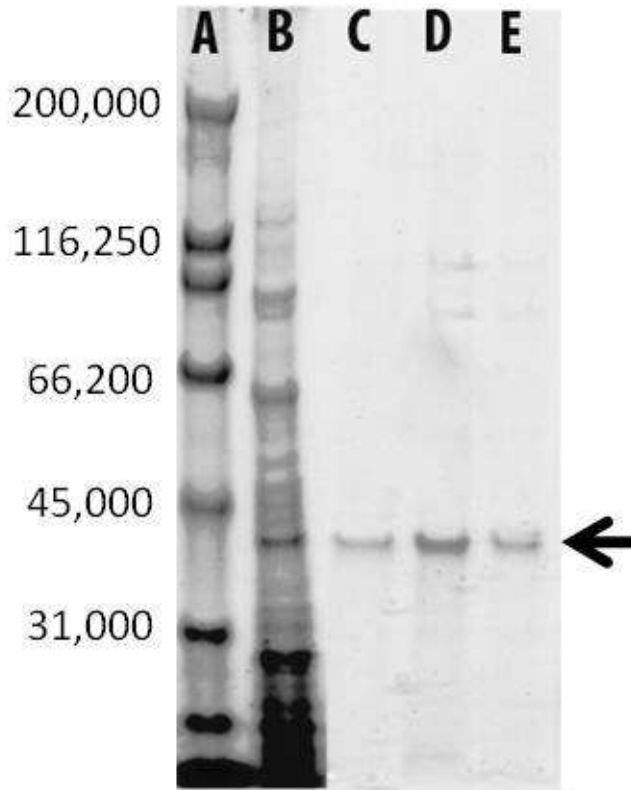


Figure S4: Results of SDS-PAGE of the component protein of Ss-vlp. Samples were fractionated by SDS-PAGE on a 12.5% polyacrylamide gel and stained with Coomassie Brilliant Blue. Lane A, Prestained molecular mass markers (Da); lane B, lysate from a culture of SsP2; and lanes C, D, E are fractions collected across the SEC peak in observed for purified Ss-vlp. The band of Ss-vlp protein, with a molecular mass of approximately 39 kDa, is indicated by an arrow.

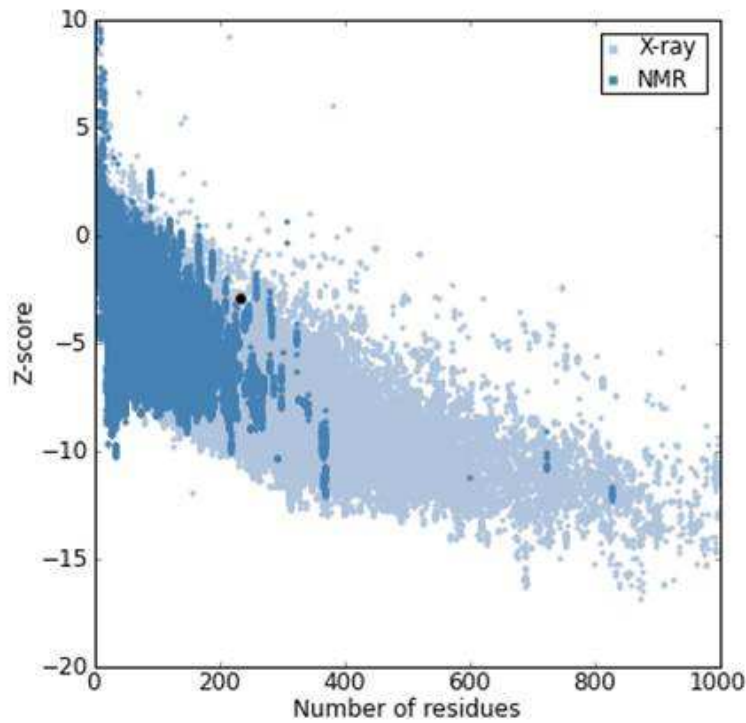


Figure S5: Quality assessment of the pseudoatomic models of the Ss-vlp (MCPs) from SsP2 provirus. Quality of the generated models as well as that of the template used for structural modelling was evaluated using ProSA-web at <https://prosa.services.came.sbg.ac.at/prosa.php>. The quality (Z) score of the modelled Ss-vlp protein is displayed in the context of the Z-scores of all experimentally determined protein structures available in the Protein Data Bank. Every dot represents a distinct structure solved by X-ray crystallography (light blue) or NMR (dark blue). 1 (Black circle) Ss-vlp model based upon PfV crystal structure (PDB id: 2E0Z, Z-score: -2.96)

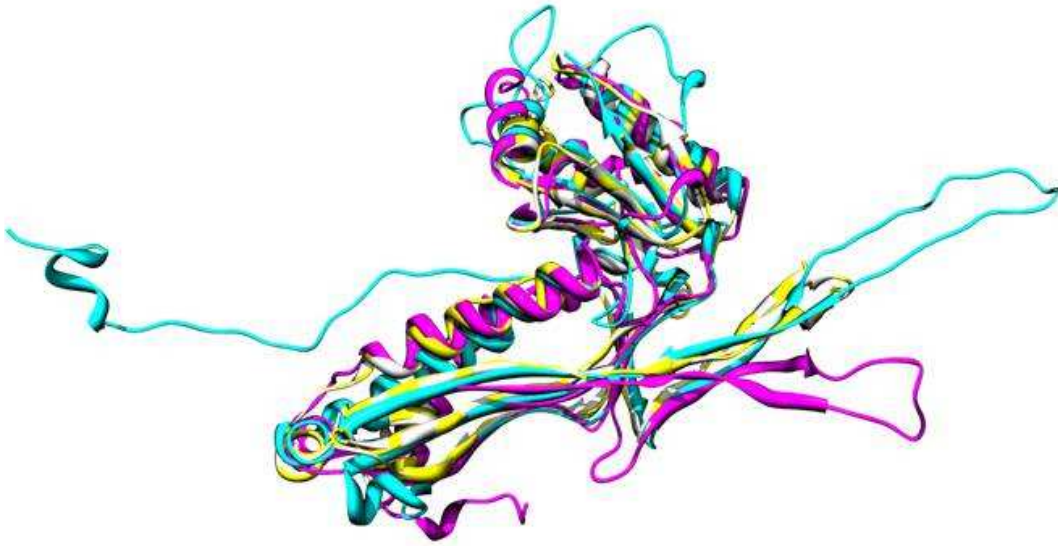


Figure S6: Structural comparison of Ss-vlp (model), *P. furiosus* Pfv, *T. Maritima* encapsulin and HK97 virus gp5. Structural alignment of Ss-vlp (yellow), Pfv (white), encapsulin (magenta) and gp5 (blue).

Figure S7:

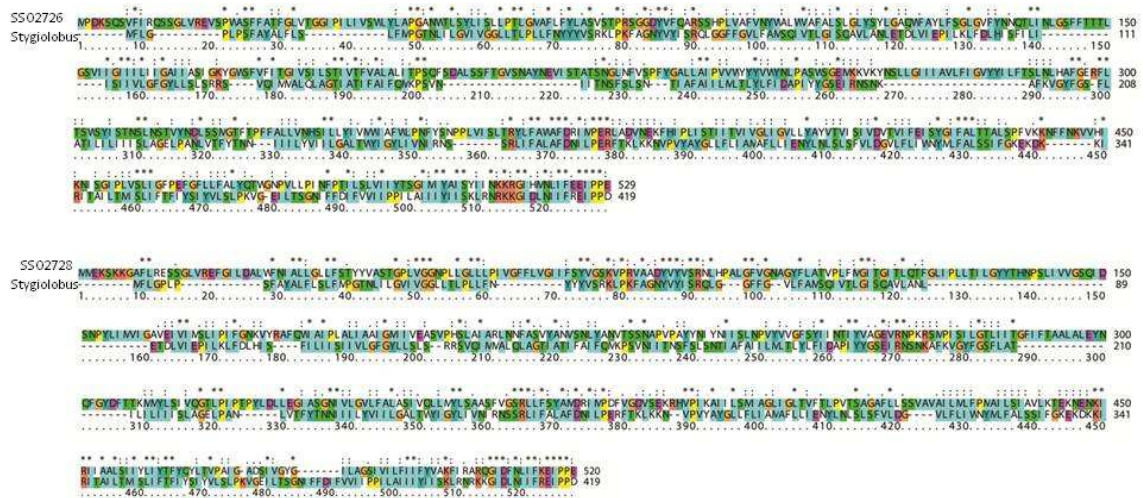


Figure S8:

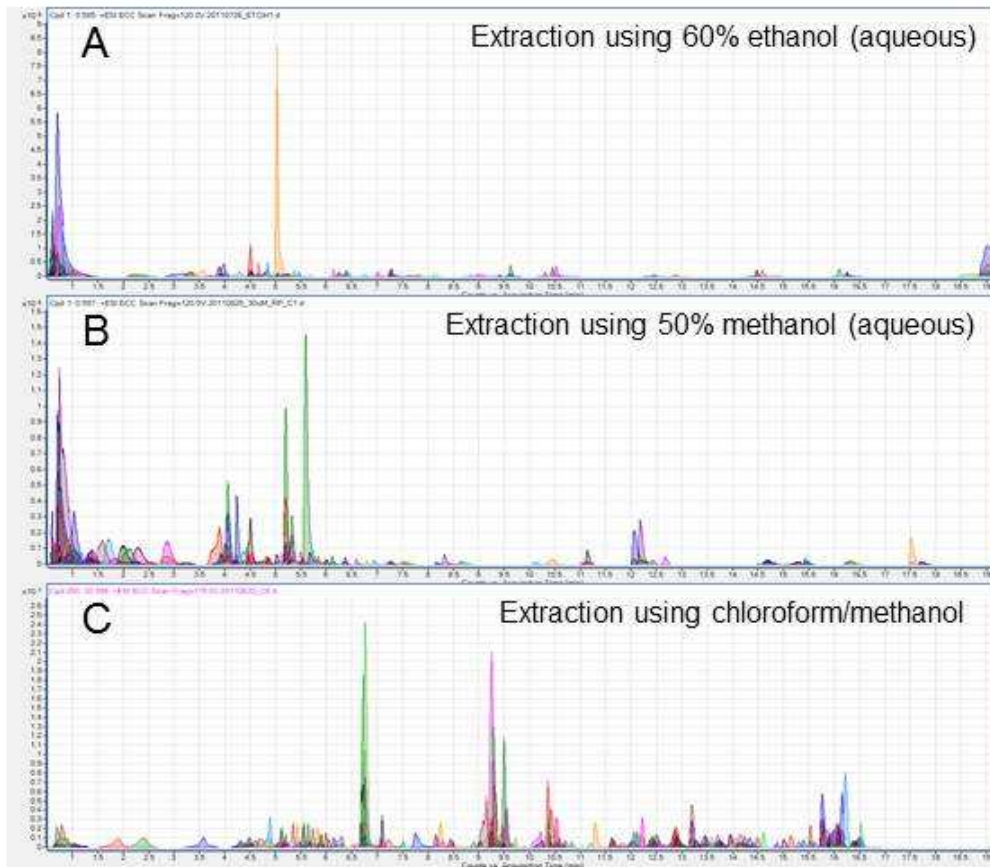


Figure S9:

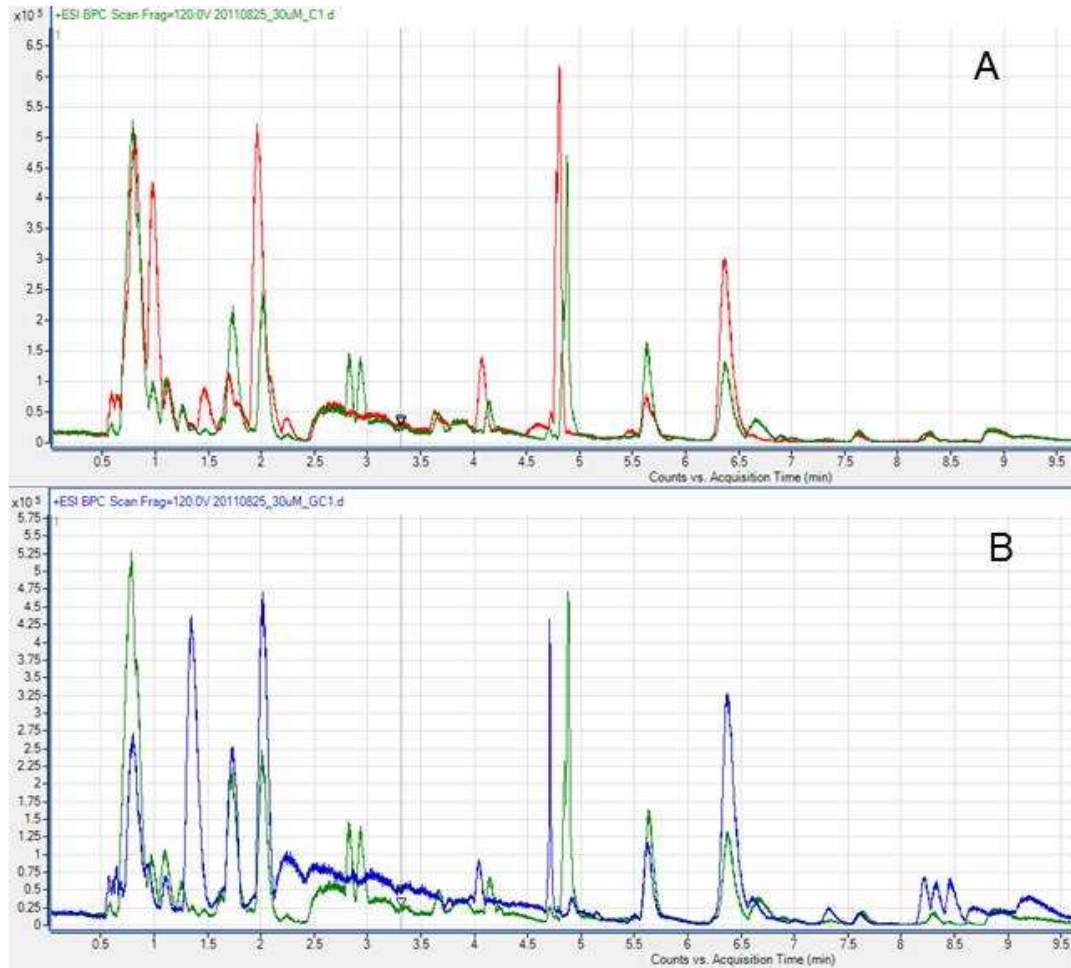


APPENDIX B

SUPPORTING INFORMATION FOR CHAPTER 3



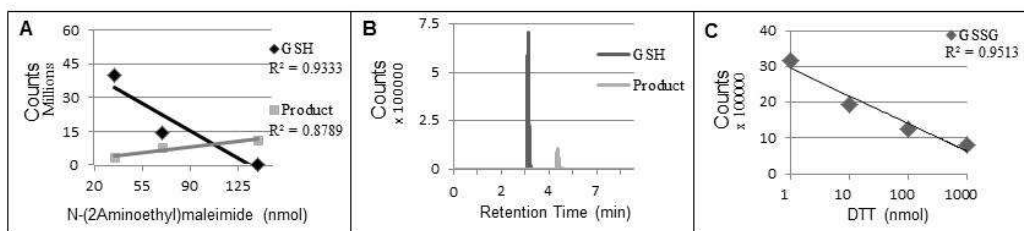
Supplemental Figure 1 Extracted ion chromatograms of 250 most abundant features: 30min linear-gradient 0-100% ACN on Phenomenex C18 column. Greater representation of hydrophobic small molecules are seen in metabolite samples extracted with a higher percentage organic solvents ($C > B > A$). Methods are highly complementary and when combined cover a significant percent of total metabolites in *S. solfataricus* (Kinetix 1.7u C18, 100A, 150mm x 2.1mm).



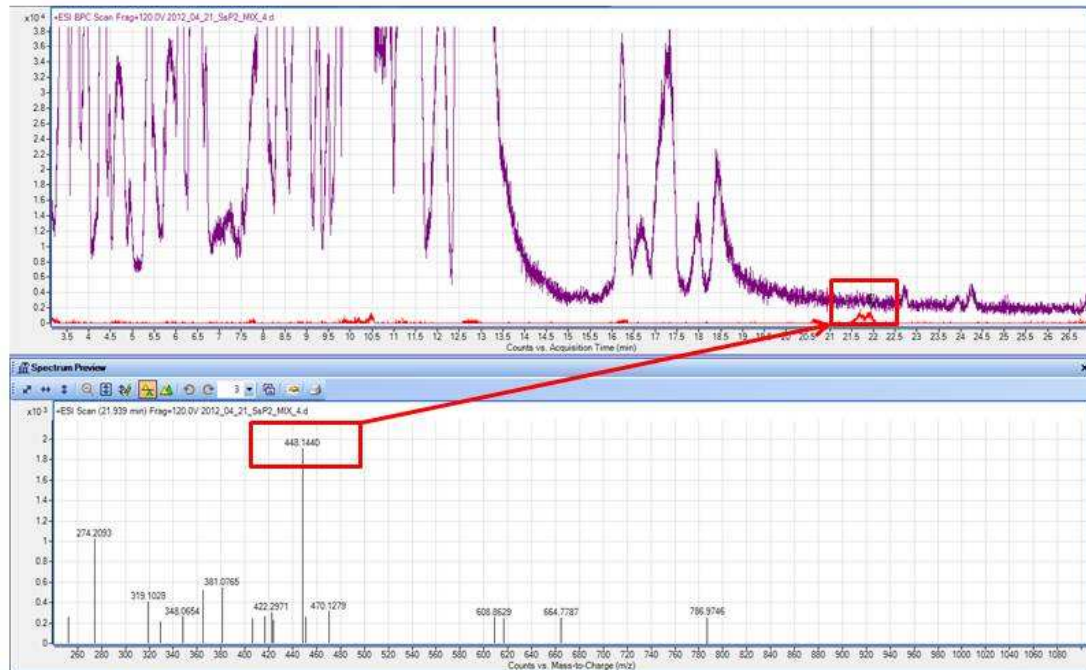
Supplemental Figure 2 Metabolite profiles depends on growth state and media. A) Base peak chromatograms (BPC) of metabolites from *S. solfataricus* harvested at stationary (red) or mid-log (green). Analysis used a HILIC (Cogent Diamond Hydride) column. B) Metabolite extraction *S. solfataricus* cells grown to the same optical density (OD) in media 1 (green), and media 2 with 0.1% glucose (blue)



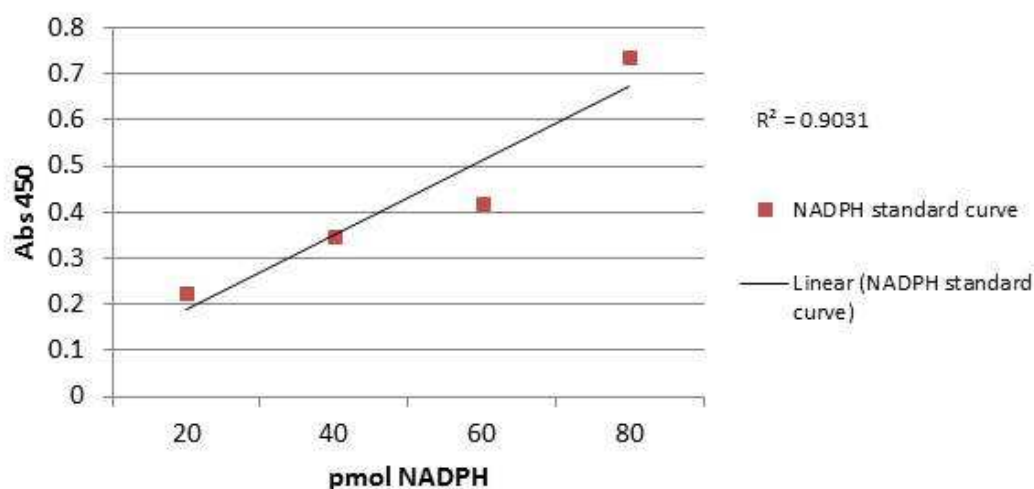
Supplemental Figure 3 Oxidized glutathione within metabolite extracts from *Ignicoccus hospitalis*. (Top) extracted ion chromatogram of 613.15 (GSSG) from metabolite extract (green), with retention time comparison with GSSG standard (blue). GSH is present in the unlabeled state, but near the level of detection. (Bottom) Mass spectrum from the elution peak at 4 minutes showing the expected m/z for GSSG.



Supplemental Figure 4 Mass spectrometry based detection and chemical modification of GSH and GSSG. (A) Labeling of Glutathione (GSH) with N-(2Aminoethyl)maleimide causes product formation. Glutathione (GSH) [m/z : 308] extracted ion chromatogram (Black); Product [m/z : 448] extracted ion chromatogram (Gray); (B) Labeling of GSH with N-(2Aminoethyl)maleimide results in product formation, as amount of N-(2Aminoethyl)maleimide is increased, product amount increases and GSH decreases linearly. (C) GSSG reduction in *S. solfataricus* metabolite extracts treated with dithiothreitol (DTT), x-scale is exponential.



Supplemental Figure 5 Labeling of reduced glutathione with N-(2Aminoethyl)maleimide within metabolite extracts from *S. solfataricus*. (Top) Base peak chromatogram of metabolite extract (purple). Extracted ion chromatogram (EIC) for m/z 448.144 corresponding to reduced glutathione (GSH) + N-(2Aminoethyl)maleimide conjugate (product of labeling reaction). GSH is present in the unlabeled state, but near the level of detection. (Bottom) Mass spectrum from the elution peak at 21.8 minutes showing the expected m/z for labeled GSH. Retention time was confirmed with a standard.



<i>S. solfataricus</i>	NADP	NADPH
Mid-log	0.54/ 62 pmol	below detection
Stationary	0.452/ 50 pmol	below detection

Supplemental Figure 6 (Top) NADPH standard curve. (Bottom) Absorbance for correlation of cells extracted at mid-log and stationary phase to determine NADPH levels according to protocol (Sigma). The amount of reduced NADPH was below the threshold of detection of the kit.

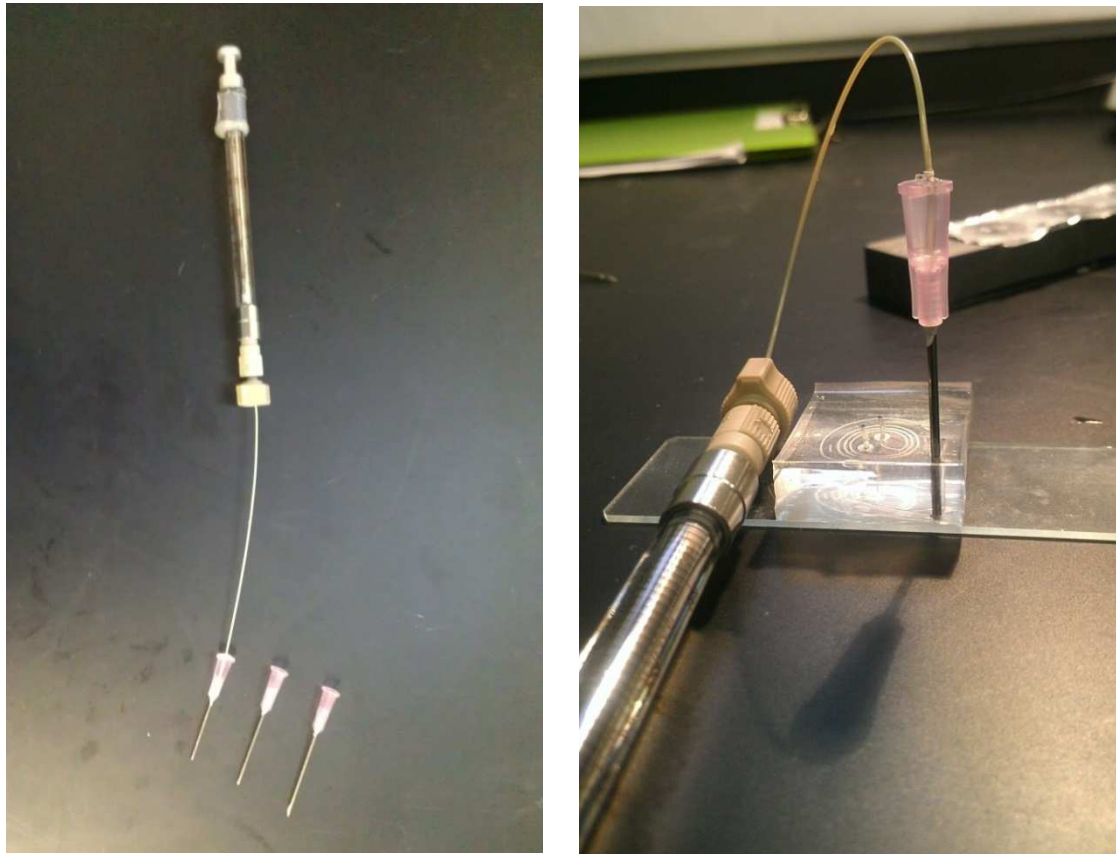
Table S1: Putative identifications for small molecule thiols labeled in *E. coli* extracts spiked with 2Aminoethylmaleimide

Name	Mass	Fold Change	P-value
Glutathione	307.0838	18.79601457	0.048054
L-Cys-Gly	178.0412	inf*	0.039185
Glu Cys Ala	321.0995	42.34292754	0.010958
Cys Met Gly	309.0817	248.9864284	0.042244

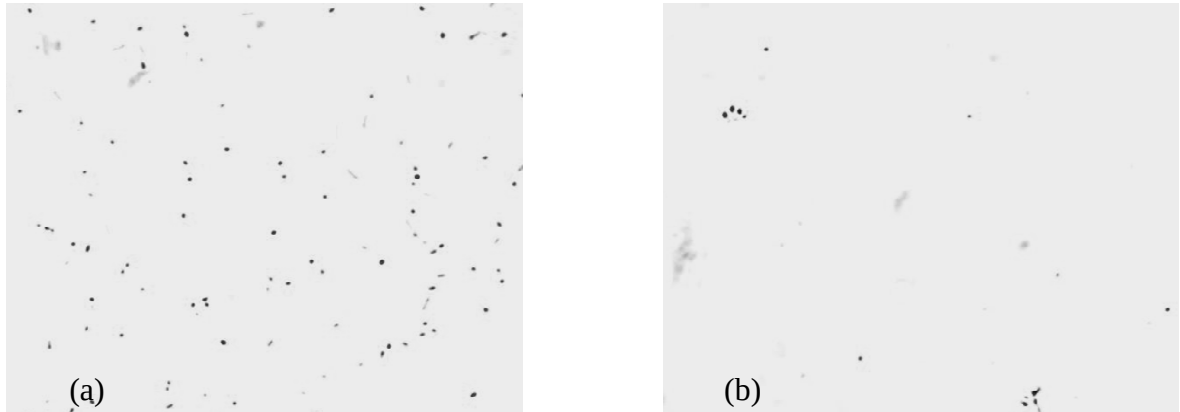
Inf* designates labeling dropped molecule below threshold of detection (Kinetix 1.7u C18, 100A, 150mm x 2.1mm).

APPENDIX C

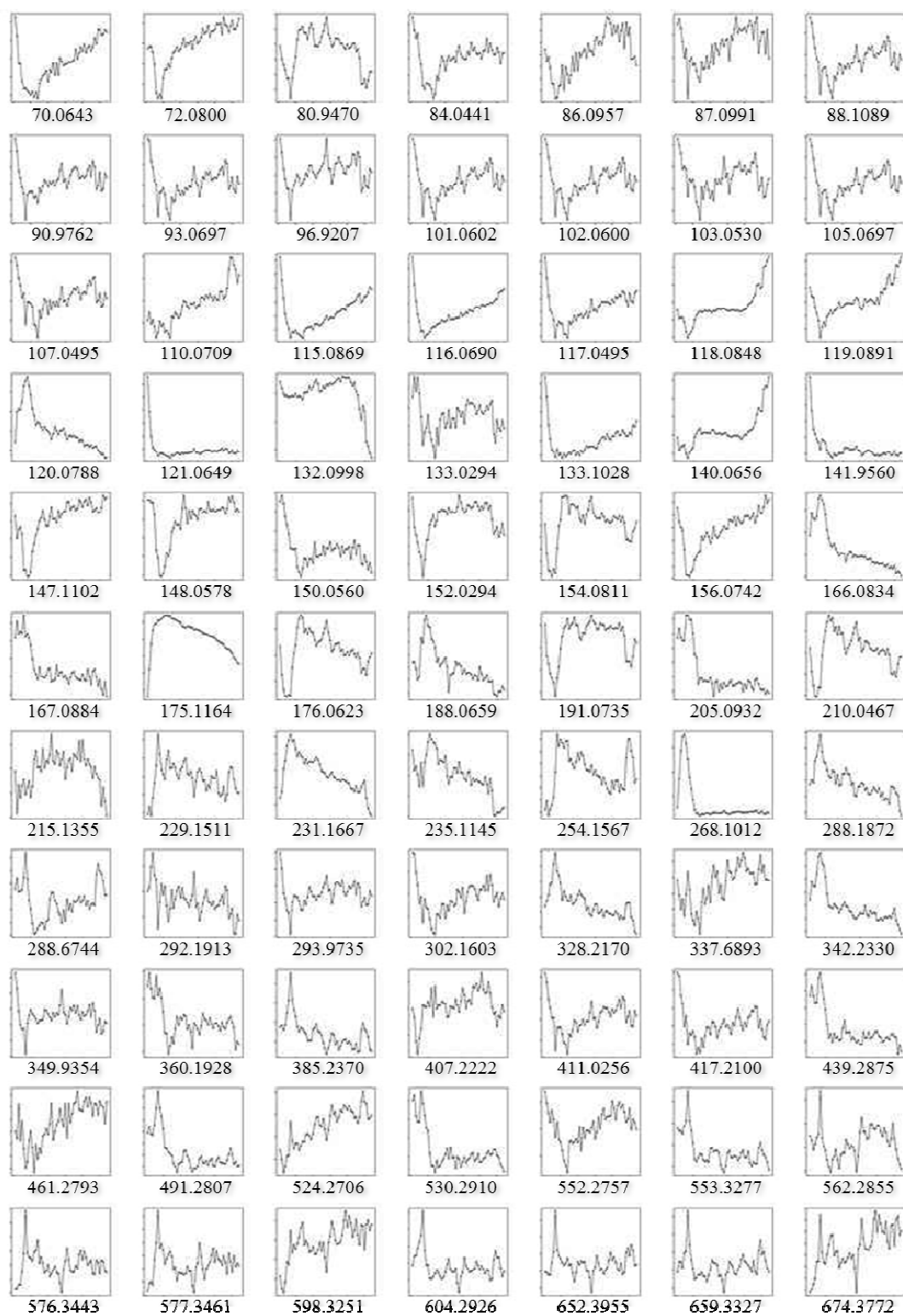
SUPPORTING INFORMATION FOR CHAPTER 5



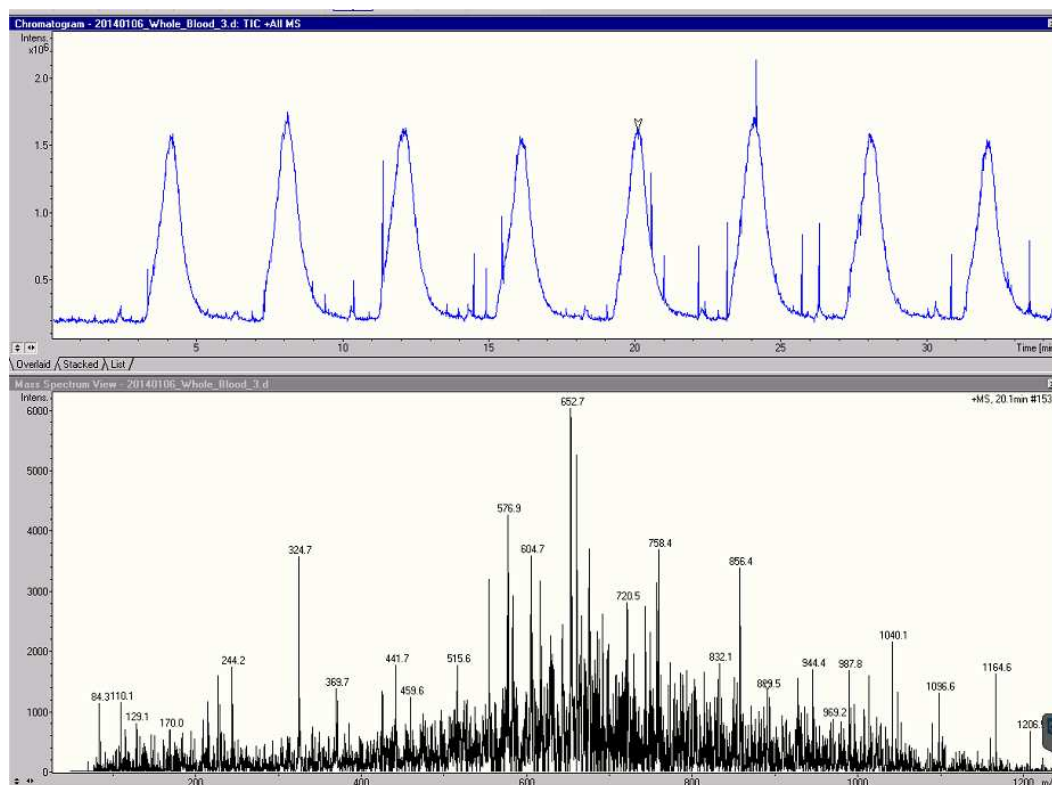
Supplementary figure 1: Custom made macro to micro interfacing using 18 Ga needle and 1/32 inch PEEK tubing. (a) Tip of 18 Ga needle was ground down flat and PEEK tubing was threaded into needle. PDMS was cured into top of needle to create an airtight seal. (b) Needle was pushed into ports in PDMS chip for easy interfacing to syringe pumps and MS.



Supplementary figure 2 Lysis of *E. coli* cell in the presence of acetonitrile. (a) Cells viewed through phase contrast microscope (1000x). (b) After exposed to ACN



Supplementary figure 3: Normalized abundance of 77 small molecules over time (~ 6 hrs, 54 timepoints).



Supplementary figure 3: Extraction of metabolites from ovine whole blood. TIC pulses (a) and mass spectra (b)