Protein stability in a proteomic perspective

Vesna Bozanic

Dissertation presented to obtain the Ph.D degree in biochemistry

Instituto de Tecnologia Química e Biológica | Universidade Nova de Lisboa

Oeiras, January, 2013

INSTITUTO
DE TECNOLOGIA
QUÍMICA E BIOLÓGICA
/UNL

Knowledge Creation

From left to right: **Doutor José Bártholo Pereira Leal** (Instituto Gulbenkian de Ciência da Fundação Calouste Gulbenkian, Lisboa), **Doutor Rune Matthiesen** (Instituto de Patologia e Imunologia Molecular - Universidade do Porto), **Doutor Cláudio Emanuel Moreira Gomes** (Instituto de Tecnologia Química e Biológica – Universidade Nova de Lisboa, ***Supervisor***), **Doutor Manuel António da Silva Santos** (Departamento de Biologia da Universidade de Aveiro), **Doutora Claudina Amélia Marques Rodrigues-Pousada** (Instituto de Tecnologia Química e Biológica da Universidade Nova de Lisboa), **Doutora Vesna Bozanic, Doutor Paulo José Garcia de Lemos Trigueiros de Martel** (Faculdade de Ciências e Tecnologia da Universidade do Algarve).

Fundação para Ciência e Tecnologia funded PhD grant SFRH/BD/18746/2004 for Vesna Bozanic and research project grant POCTI/BIO/58465 (PI: Cláudio M. Gomes).

Foreword

This thesis is the result of research work done at Protein Biochemistry Folding and Stability Laboratory at Instituto de Tecnologia Química e Biológica (ITQB), Universidade Nova de Lisboa (UNL), Portugal, under the supervision of Professor Doctor Cláudio M. Gomes.

The studies here reported were performed during the term of a four year PhD fellowship from Fundação para a Ciência e Tecnologia (FCT), from February 2005 to February 2009.

The thesis comprises six chapters. The first chapter introduces the reader to the life at high temperatures and addresses mechanisms through which proteins acquire enhanced stability.

The second chapter addresses methodologies that are used in proteomics studies of protein stability.

Chapter Three outlines the results regarding a proteomic study of selected hyperstable cytosolic proteins originating from the thermophile *Sulfurisphaera sp*.

Chapter Four presents results regarding a comparative study of thermostable proteins selected from the cytosolic proteome of the thermophile *Sulfolobus solfataricus* and the mesophile *Escherichia coli*.

Chapter Five discusses the subset of results regarding hyperstable cytosolic proteome from *Escherichia coli* discussing relationships between protein thermostability and solubility.

The last chapter of the thesis presents in brief concluding remarks regarding the work presented.

Acknowledgments

I would like to express my gratitude to the following people and institutions that have contributed to my PhD thesis, among them especially:

- My supervisor, Professor Doctor Cláudio M. Gomes
- All my colleagues, past and present members of the Protein Biochemistry Folding and Stability Laboratory: Dr. Bárbara Henriques, Catarina Silva, Dr. Hugo Botelho, Dr. João Rodrigues, Dr. Patrícia Faísca, Dr. Raquel Correia, Dr. Sónia S. Leal.
- Instituto de Tecnologia Quimica e Biológica (ITQB) and Prof. Dr. L.P. Rebelo

Collaborating laboratories:

- Prof. Dr. Phillip Wright, head of Department of Chemical and Biological Engineering, University of Sheffield, UK with Dr. Trong Khoa Pham, for enabeling iTRAQ experiments and MS data analysis.
- Dr. José P. Leal, head of Computational Genomics Laboratory, Instituto Gulbenkian de Ciência, with Renato Alves, for providing computational algorithms.

Funding of PhD grant and research project:

- Fundação para Ciência e Tecnologia for funding my PhD grant SFRH/BD/18746/2004 and research project grant POCTI/BIO/58465 (PI: Cláudio M. Gomes).

My deepest gratitude goes to:

- My family and friends

Publications

The work presented in this thesis is based on the following publications in international peer reviewed journals:

Vesna Prosinecki, Hugo M. Botelho, Simona Francese, Guido Mastrobuoni, Gloriano Moneti, Tim Urich, Arnulf Kletzin and Cláudio M. Gomes, A Proteomic Approach toward the Selection of Proteins with Enhanced Intrinsic Conformational Stability, Journal of Proteome Research**, 2006,** 5 (10): 2720-2726

Vesna Prosinecki, Patrícia F.N. Faísca and Cláudio M. Gomes, Conformational States and Protein Stability in a Proteomic Perspective, Current Proteomics, 2007, 4 (1): 44-52

Additional publications resulting from the work during the term of PhD grant from Fundação para a Ciência e Tecnologia (FCT):

Hugo M. Botelho, Sónia S. Leal, Andreas Veith, **Vesna Prosinecki**, Christian Bauer, Renate Frohlich, Arnulf Kletzin, Cláudio M. Gomes Role of a novel disulfide bridge within the all-beta fold of soluble Rieske proteins**,** J Biol Inorg Chem, 2010, 15(2):271-81

João V. Rodrigues, **Vesna Prosinecki**, Isabel M. Marrucho,, Luís P. Rebelo, Cláudio M. Gomes, Protein stability in an ionic liquid milieu: on the use of differential scanning fluorimetry**,** Phys. Chem. Chem. Phys., 2011, 13(30):13614-6

Abstract

This work involved the identification and analysis of the properties of the most stable proteins present within proteomes, aiming at obtaining a general perspective of the factors that determine protein stability. As models we have focused on ensembles of proteins with high intrinsic stability, and for this purpose we have studied proteomes from the hyperthermophilic archaeon *Sulfolobus solfataricus* and *Sulfurisphaera sp*., whose properties were compared to those of the mesophilic bacterium *Escherichia coli.*

To carry out this study, we have implemented a novel approach aimed at profiling a soluble proteome for its most intrinsically stable proteins. For this purpose the hyperthermophilic archaeon *Sulfurispharea sp*., which is able to grow between 70-97°C, was used as a model organism. We have thermally and chemically perturbed the cytosolic proteome as a function of time (up to 96h incubation at 90°C), and proceeded with analysis of the remaining proteins by combining one and two dimensional gel electrophoresis, liquid chromatography fractionation, protein identification by N-terminal sequencing, and mass spectrometry methods. A total of 14 proteins with enhanced stabilities which are involved in key cellular processes such as detoxification, nucleic acid processing and energy metabolism were identified. We demonstrate that these proteins are biologically active after extensive thermal treatment of the proteome. This method has thus illustrated an experimental approach aimed at mining a proteome for hyperstable proteins, a valuable tool for target selection in protein stability and structural studies, and biotechnological applications.

The hyperthemophilic organism *Sulfolobus solfataricus* and the mesophilic bacterium *Escherichia coli* were selected as models for furtherer investigation with regard to the profiling of proteins at a proteomic scale according to their stabilities. The previously established thermal perturbation methodology was employed, but now also in combination with iTRAQ mass spectrometry analysis. This has allowed identification and quantification of the relative variations of individual proteins along the thermal perturbation process. This has resulted in the

definition of three groups of proteins (around 300 distinct proteins in each organism) corresponding to the subset of proteins with above average stability (labeled 'survivors'), unchanged stability, and less stable than the average. These sequences were investigated using bioinformatics tools, in order to determine relationships between thermostability, physicochemical properties, structural folds, amino acid type and biological function.

We concluded that *per se* the prevalence of certain types of amino acids is not essential to make a protein more stable and that SCOP folds are also not strongly biased in respect to thermostability. The group of thermostable proteins was slightly enriched in smaller proteins (<50kDa), with slightly negative GRAVY scores and higher aliphatic indexes. Regarding COG functional categories, the identified sequences with increased stability and solubility belonged to the following categories: Information storage and processing group (27%), cellular processes group (30%) and metabolism group (40%). This could suggest that these particular processes have evolved so as to preserve stable folds. In any case our results show that enhanced thermal stability results from a combination of properties, and not from a single exclusive factor.

Another interesting finding is that enhanced stability seems to be correlated with increased solubility, as shown from the comparison of our results with those obtained in an independent study that compared the solubility/aggregation ratio for soluble *Escherichia coli* proteins expressed in a cell-free and chaperone-free system. Indeed, analysis of the solubility/aggregation profile of the superstable cytosolic proteins that survived harsh thermal treatment shows that these are mostly overlapping with those which rank soluble. This denotes that protein stability and solubility are intertwined properties grounded in comparable physical principles, as selection for stability yields increased solubility as a read-out.

Altogether, the experimental method developed in this work proved to be an excellent tool in mining for proteins with high stability and low propensity to aggregate. Having established the proof of concept of its applicability on the studied proteomes, this methodology can be easily

applied to any proteome or complex protein mixture of soluble proteins. Based on this, further development of experimental, theoretical and computational approaches on entire ensembles of proteins from a given organism will result in a better understanding of the physical principles underlying protein folding, stability and solubility.

Resumo

O trabalho apresentado nesta tese envolveu a identificação e análise das propriedades de proteínas com estabilidade elevada no contexto do proteoma onde se inserem, visando a obtenção de uma perspetiva geral acerca dos determinantes da estabilidade proteica. Como modelos selecionamos conjuntos de proteínas com elevada estabilidade intrínseca, pelo que se estudaram os proteomas dos Archaea hipertermófilos *Sulfolobus solfataricus* e *Sulfurisphaera sp*., cujas propriedades foram comparadas com as da bactéria mesófila *Escherichia coli.*

A implementação deste estudo envolveu o desenvolvimento de uma nova metodologia que permitiu selecionar as proteínas constituintes de um proteoma de acordo com a sua estabilidade relativa. Para o efeito estudou-se o hipertermófilo Sulfurisphaera sp., que cresce entre 70-97°C, como modelo. O protocolo desenvolvido consistiu na perturbação química e térmica do proteoma citosólico em função do tempo (até 96h de incubação a 90°C), seguido da análise das proteínas remanescentes usando técnicas cromatográficas, eletroforese de proteinas (1D e 2D), identificação por sequenciação do N-terminal e espectrometria de massa.

Um conjunto de 14 proteínas foi identificado neste estudo piloto, estando estas sobretudo implicadas em processos celulares essenciais como destoxificação, processamento de ácidos nucleicos e metabolismo energético. Verificou-se igualmente que estas proteínas permaneciam biologicamente ativas após o processo de perturbação térmica. Este método ilustra assim um modo de pesquisar proteínas híper estáveis, uma valiosa ferramenta na seleção de alvos para estudos estruturais e de estabilidade proteica, assim como aplicações biotecnológicas.

Numa segunda fase, o aprofundar destes estudos envolveu o recurso aos proteomas do hipertermófilo *Sulfolobus solfataricus* e do mesófilo *Escherichia coli*. O método de perturbação térmica foi aplicado, mas desta vez complementado com análise iTRAQ por espetrometria de massa, o que permitiu a identificação e quantificação de proteínas

individualmente ao longo do processo de perturbação. Daqui resultou a definição de três grupos de proteínas (cerca de 300 proteínas distintas em cada organismo) correspondendo a subconjuntos de proteínas com 1) estabilidade acima da média (designados 'sobreviventes'), 2) inalteradas e 3) com estabilidade decrescida. Estas sequências foram analisadas recorrendo a ferramentas de bioinformática de modo a estabelecer correlações entre termo-estabilidade, características físico-químicas, tipos de estrutura, amino ácidos e função biológica.

Concluiu-se que *per se*, a prevalência de um certo tipo de amino ácido(s) não determina a estabilidade de uma proteína e que os tipos de estrutura de acordo com a classificação SCOP não denotam igualmente uma preponderância. O grupo de proteínas hiperestáveis denotou um ligeiro enriquecimento em proteínas relativamente pequenas (< 50 kDa), com scores GRAVY negativos e índices alifáticos mais elevados. No que diz respeito às categorias funcionais COG, a maioria das proteínas pertence a um dos grupos seguintes: grupo armazenamento e processamento de informação, processos celulares e metabolismo. Isto sugere que estes processos em particular podem ter evoluido de modo a preservar proteínas estáveis. Globalmente os resultados obtidos sugerem que a estabilidade acrescida resulta de uma combinação de factores.

Uma outra observação muito relevante foi a de uma correlação entre estabilidade e solubilidade, que resultou da comparação do rácio solubilidade/agregação determinado num estudo independente para o proteoma solúvel de *E. coli* a partir da expressão individual de cada uma das proteínas num sistema *cell free*. A análise das proteínas que resistiram ao tratamento de perturbação térmica de acordo com este índice revelou que estas são predominantemente solúveis. Esta observação indica que estabilidade e solubilidade proteica são propriedades conexas e fundamentadas nos mesmos princípios físicos e químicos, na medida em que o resultado da seleção para o factor 'estabilidade' é o factor 'solubilidade'.

Globalmente, a metodologia experimental desenvolvida neste trabalho revelou-se uma excelente ferramenta para pesquisar proteínas com elevada estabilidade e baixa propensão para agregar. Tendo estabelecido

o princípio da aplicabilidade deste método aos proteomas estudados, a metodologia pode ser facilmente expandida para o estudo de qualquer outro proteoma ou mistura complexa de proteínas. A partir destes resultados, espera-se que o desenvolvimento de métodos experimentais, teóricos e computacionais com base nestes princípios possa levar a uma melhor compreensão dos princípios físicos subjacentes ao folding, estabilidade e solubilidade proteica.

Abbreviations

1DE	One dimensional
2DE	Two dimensional
Å	Angstrom
AI	Aliphatic index
ASCA	Complete Set of E.coli K-12 ORF Archive
ATP	Adenosine triphosphate
BCP	Bacterioferritin co-migratory protein
Bis-ANS	4,4'-bis(1-anilinonaphthalene 8-sulfonate)
CATH	Hierarchical domain classification of protein structures
CD	Circular Dichroism
CHAPS	3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonate
CID	Collision Induced Dissociation
COG	Clusters of Orthologous Groups
Da	Dalton
DLS	Dynamic Light Scattering
DNA	Deoxyribonucleic acid
DTT	1,4-Dithiothreitol
E. Coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
eg.	example given (for example)
et al.	And others
Fd	Ferredoxin
GDP	Guanosine diphosphate
GRAVY	Grand average of hydropathicity index
GTP	Guanosine triphosphate
GuHCl	Guanidine hydrochloride
IEF	Isoelectric focusing
IPG	Imobilizad pH gradient
IR	Infrared spectroscopy
iTRAQ	Isobaric Tags for Relative and Absolute Quantitation
M	Molar
MALDI	Matrix Assisted Laser Desorption Ionization
mRNA	Messenger RNA
MS	Mass Spectrometry
MW	Molecular weight

NCBI	National Center for Biotechnology Information
NL	Non linear
NMR	Nuclear Magnetic Resonance
OGT	Optimal growth temperature
ORF	Open reading frames
PAGE	Polyacrilamide gel electrophoresis
PDB	Protein Data Bank
pI	Isoelectric point
PMSF	Phenylmethylsulfonyl fluoride
Prx	Peroxiredoxin
RNA	Ribonucleic acid
ROS	Reactive oxygen species
rRNA	Ribosomal ribonucleic acid
S. solfataricus	*Sulfolobus solfataricus*
SCOP	Structural Classification of Proteins
SDS	Sodium dodecyl sulfate
SOD	Superoxide dismutase
TCA	Trichloroacetic acid
TF	Transcription factor
TFA	Trifluoroacetic acid
Tm	Midpoint transition temperature
Topt	Optimal temperature
Tris-HCL	Tris(hydroxymethyl)aminomethane hydrochloride
UV	Ultraviolet
vs.	*Versus*

Aminoacids

A	Ala	Alanine
C	Cys	Cysteine
D	Asp	Aspartate
E	Glu	Glutamate
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophane
Y	Tyr	Tyrosine

Table of contents

This chapter was partially published in:

Vesna Prosinecki, Patrícia F.N. Faísca and Cláudio M. Gomes, Conformational States and Protein Stability in a Proteomic Perspective, Current Proteomics, Vol. 4 Issue 1, 2007, 44-52

Chapter 1

INTRODUCTION

1.1.Biodiversity of thermophiles

Thermostable organisms and their proteins have been subject of research during the last decades due to their various unique properties. Interest in thermophiles and how their proteins manage to function at elevated temperatures started in 1960's by the pioneering work of Brock and his colleagues (1) and continues up to date. Even nowadays, elevated interest in those remarkable organisms still continues, aiming at exploring mechanisms of survival at the environmental extremes, offering valuable data in knowing bases of protein stability.

Based on their optimal growth temperatures (OGT) organisms are divided into three main groups, *i.e*. psychrophiles with OGT below 20°C, mesophiles that optimally grow at moderate temperatures from 20°C up to 55°C, and thermophiles that thrive in high temperatures, above 55°C. Only few eukaryotes are known to grow above this temperature, but some fungi grow in the temperature range 50 up to 60°C (2). In 1992 Kristjansson and Stetter (3) suggested a further division of the thermophiles and a hyperthermophile boundary by growth at and above 80°C. The Tree of Life that is 16S rRNA-based phylogenetic tree exhibits three domains, the Bacteria, Archaea and Eukarya. The use of ribosomal RNA sequences has led to the recognition of a group of prokaryotes, the Archaea that are lacking a nuclear membrane and possessing a single circular chromosome, while possessing several molecular properties with similarity to the eukaryotes such as transcription signals, transcription factors, chaperones, and histones. Archaea are phylogenetically distinct from both Bacteria and Eukarya and rich in thermophilic and hyperthermophilic species. Members of the deepest and shortest lineages exhibit the highest growth temperatures (Fig. 1.1 (5)).

Hyperthermophiles are well adapted to their biotopes, are able to grow not just at high temperatures but often at extremes of pH, redox potential, and salinity. They are found in vast range of natural and artificial water-containing hot environments. Marine biotopes' hyperthermophiles are adapted to the high salinity of sea water. Terrestrial ones usually require low salinity. Most hyperthermophiles, with some exceptions, are strict anaerobes. Depending on the energy

sources available, they show great versatility: members of the same genera and even the same strains may be able to use different electron donors and acceptors. In addition, several hyperthermophilic Archaea are facultative or obligate heterotrophs able to use organic compounds as energy and carbon sources (4, 5). Most thermophilic bacteria characterized today grow below 80°C hyperthermophilic boundary with OGT up to 50°C with some exceptions, such as *Thermotoga* and *Aquifex* (4) while hyperthermophilic species are highly dominated by the Archaea. Currently, the most thermophilic organism known is *Pyrolobus fumarii* that grows in the temperature range of 90 to 113°C. Regarding thermophiles, the bacteria *Aquifex pyrophilus* and *Thermotoga maritima* exhibit the highest growth temperatures of 95°C (4). Within the Archaea, the organisms with the highest growth temperatures, between 102 and 113°C, are found within the *Crenarchaeota* and the *Euryarchaeota* (6). They are members of the crenarchaeal genera *Pyrolobus, Pyrodictium, Hyperthermus, Pyrobaculum, Igneococcus* and *Stetteria* and the euryarchaeal genera *Methanopyrus* and *Pyrococcus*. The upper temperature at which life is possible is still unknown, but it is probably not much above 113°C. Above 110°C most of the biological molecules become highly unstable, ATP is spontaneously hydrolyzed in aqueous solution at temperatures below 140°C, and hydrophobic interactions weaken significantly (7).

At present, vast number of species of hyperthermophilic Archaea and Bacteria are known which had been isolated from different terrestrial and marine thermal areas in the world. Hyperthermophiles are very divergent, in terms of both their phylogeny and physiological properties and are grouped into 34 genera and 10 orders (5). Due to the fact that hyperthermophiles belong to two phylogenetically distinct domains of life, the Bacteria and Archaea, the strategies of molecular mechanisms of heat adaptation may be quite different depending on the phylogenetic position of the corresponding organism.

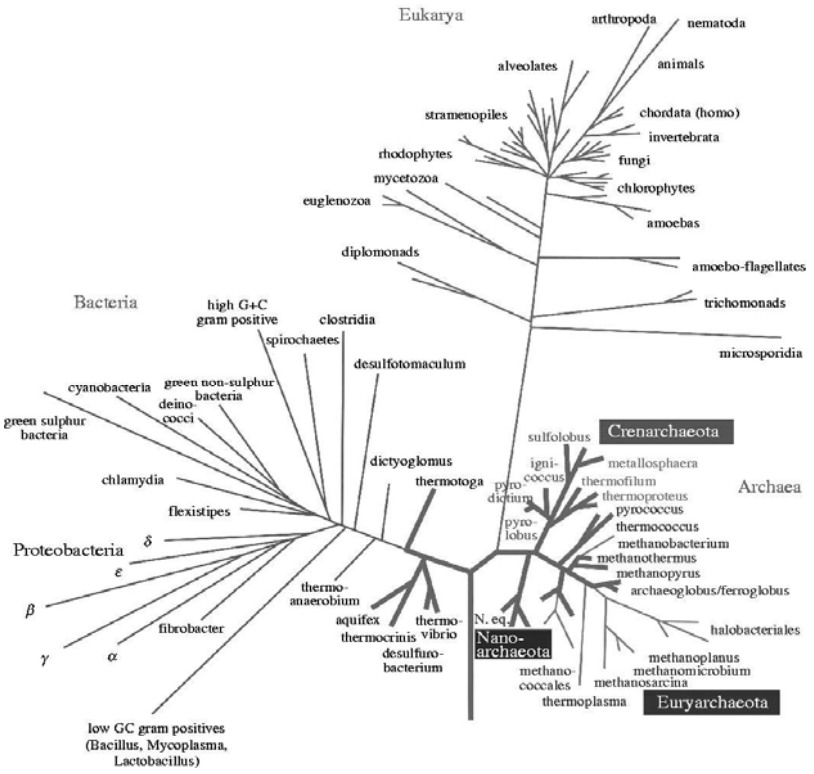


Fig. 1.1. Small subunit ribosomal RNA-based phylogenetic tree. The thick lineages represent hyperthermophiles (5).

1.2.Model organisms

The focus of this research work and thesis has been on proteins with enhanced stability properties originating from two distinct groups of organisms: Thermophiles *Sulfolobus solfataricus* and *Sulfurisphaera sp*. from the archaeal order *Sulfolobale*s, and the mesophilic bacteria *Escherichia coli*. These models are briefly presented.

1.2.1. Thermophilic model organisms: *Sulfolobales*

Order of *Sulfolobales* belongs to category of sulfur dependent Archaea. They are of aerobic or facultatively aerobic, chemolithotrophic cocci and extreme thermoacidophiles. They lack peptidoglycan in their cell walls. *Sulfolobales* grow in terrestrial volcanic hot springs with optimum growth occurring at pH 2-3 and a temperature of 75-80°C. Typically they obtain

energy for growth by the oxidation of sulfur to sulfuric acid. *Sulfolobales* are generally less thermophilic than *Thermoproteales* and *Thermococcales*, with only species of *Acidianus* being able to grow at or above 90°C. In addition, they mainly inhabit continental spring waters rich in sulfur, although some species are also found near shallow marine volcanic vents. Species of *Acidianus* and *Desulfurolobus* also grow under anaerobic conditions by the reduction of sulfur to H_2S using H2 as the electron donor. *Stygiolobus* is unique among the *Sulfolobales* as it does not grow under aerobic condition.

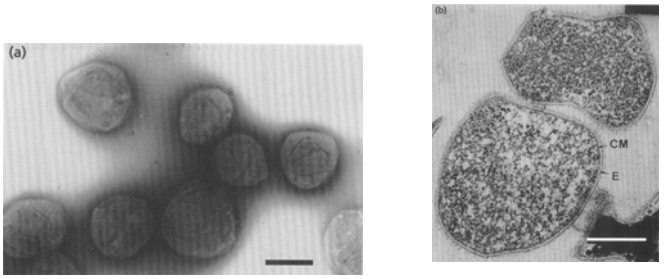


Fig. 1.2 *Sulfurisphaera sp.* Electron micrographs of strain TA-lT. (a) Negative staining; bar, 1 prn. (b) Thin section. Cell membrane (CM) and envelope (E) are indicated; bar, 0.5 pm. (8)

Sulfolobus solfataricus* and *Sulfurisphaera sp.

Sulfurisphaera is facultatively anaerobic, thermophilic, Gram-negative crenarchaeon that occur in acidic solfataric fields. The organism grows under the temperature range of 63-92°C with the optimum temperature at 84°C, and under the pH range of pH 1.0 and 5.0 with optimum pH 2.0. It forms colonies that are smooth, roundly convex, and slightly yellow. The strains of Sulfurisphaera ohwakuensis were isolated from multiple locations in the acidic hot springs in Ohwaku Valley, Hakone, Japan. (8) (Fig. 1.2. (8)). Taxonomic hierarchy of Order *Sulfolobales* is presented in Scheme 1.1.

Domain	*Archaea* C.R. Woese et al., 1990
Phylum	*Crenarchaeota* G.M. Garrity & J.G. Holt, 2001
Class	*Thermoprotei* A.L. Reysenbach, 2002
Order	*Sulfolobales* Stetter, 1989
Family	*Sulfolobaceae* Stetter, 1989
Genus	*Sulfolobus* Brock et al., 1972
	Acidianus Segerer et al., 1986
	Metallosphaera Huber et al., 1989
	Stygiolobus Segerer et al., 1991
	Sulfurisphaera Kurosawa et al.1998
	Sulfurococcus Golovacheva et al., 1995

Scheme 1.1. Taxonomic hierarchy of Order *Sulfolobales* (http://www.taxonomy.nl/taxonomicon/)

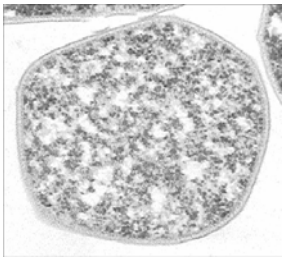


Fig. 1.3. Microscopy image of S. solfataricus. Image D.Janckovik and W.Zillig

Sulfolobus solfataricus was discovered by Wolfram Zillig and Karl Stetter in Pisciarelli near Naples, Italy (Fig. 1.3). Up to date it is the most widely studied organism of the crenarchaeal branch of the Archaea, a model for research on archaeal mechanisms and cellular processes like DNA replication, the cell cycle, chromosomal integration, transcription, RNA processing, and translation. It is shaped as highly irregular lobed cocci which usually occur singly, has no flagella, but pilus-like and pseudopodium-like structures are often found. Strictly aerobic, its optimal growth temperature is at 87°C and pH 2 to 4, metabolizing sulfur. It produces sulfuric acid. *S. solfataricus* has been isolated worldwide from continental solfatara fields including Yellowstone National Park, Mount St. Helens, Iceland, Italy, and Russia - almost wherever there is volcanic activity. The genome of the *S. solfataricus P2* contains 2,992,245 bp on a single chromosome and encodes 2,977 proteins and many RNAs. One third of the encoded proteins have no detectable homologs in other sequenced genomes. Moreover, 40% appear to be archaeal-specific, and only 12% and 2.3% are shared exclusively with Bacteria and Eukarya, respectively (9).

1.2.2. Mesophilic model organism: *Escherichia coli*

A mesophilic bacterium Escherichia coli was discovered in 1885 by the German bacteriologist Dr. Theodor Escherich and since then has been

ithe most widely used prokaryotic system for the synthesis of heterologous proteins and model in scientific research. Many decades of research have resulted in a wealth of genetic, biochemical, and structural information that together is unparalleled in other systems (10). Taxonomic hierarchy of *Escherichia coli* is presented in Scheme 1.2.

E. coli is the member of large bacterial family *Enterobacteriaceae* that are facultatively anaerobic Gram-negative rods that live in the intestinal tracts of humans and animals in health and disease, physiologically versatile and well-adapted to variations of its characteristic habitats. The genome of *E. coli* consists of 4,639,221 bp of circular duplex DNA (11) that was sequenced in 1997.

Characterization and comparison of *E. coli* paralogous proteins and protein groups and comparison to other species allows examination of the evolutionary events surrounding protein diversification. Therefore, this organism was our obvious choice for the well known model system for mesophilic organism to compare the results originating from our study regarding archaeal proeome.

Domain	*Bacteria* Haeckel, 1894, C.R. Woese et al., 1990
Phylum	*Proteobacteria* Garrity et al., 2005
Class	*Gammaproteobacteria* Garrity et al., 2005
Order	*Enterobacteriales*
Family	*Enterobacteriaceae* Rahn, 1937, nom. cons
Genus	*Escherichia coli* Migula,1895, Castellani & Chalmers, 1919

Scheme 1.2. Taxonomic hierarchy of *Escherichia coli* (http://www.taxonomy.nl/taxonomicon/)

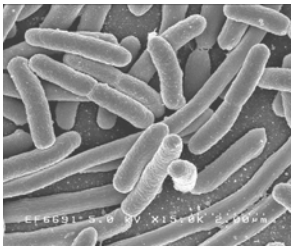


Fig. 1.4. Scanning electron micrograph of Escherichia coli, grown in culture and adhered to a cover slip. Credit: Rocky Mountain Laboratories, NIAID, National Institute of Health.

1.3.Convergent Evolution Theory of Thermal Adaptation

Dealing with the highly thermostable proteins originating from phylogeneticaly distinct organisms to start with, required the insight into possible background of origins of thermostability. Data from the literature indicate that the choice of a particular strategy depends on the evolutionary history of an organism (12):

Some hyperthermostable proteins are significantly more compact than their mesophilic homologues, without a particular interaction type causing stabilization, but a vast range of various interactions are responsible for thermostability. This stabilization strategy can be named as “structure-based”. Some other hyperthermostable proteins employ an alternative, “sequence-based” mechanism of their thermal stabilization. They do not show strong structural differences when compared to their mesophilic homologues but a small number of apparently strong interactions that are responsible for high thermal stability of these proteins. Structure-stabilized proteins come mostly from hyperthermophilic archaeal species, whereas sequence-stabilized proteins are mostly bacterial. Such differences can be attributed to the different phylogenetic background of these organisms. It is widely accepted that the archaeal environmental habitat was hot. During evolution in such kind of environment, archaeal proteins were initially designed in a hot environment in a way that its inherent structural properties enable thermal resistance with sequences able to fold and be stable in such thermostable structures. Alternatively, bacterial species that evolved later, initially as a mesophilic organism that only later recolonized a hot environment were involved in secondary thermophilic adaptation that required the enhancement of the thermostability of already existing proteins. Comparative analysis of structures and complete genomes of several hyperthermophilic archaeal organisms (e.g. *Pyrococcus furiosus*) and bacterial (12) (e.g. *Thermatoga maritima*), revealed that organisms develop diverse strategies of thermophilic adaptation by using these two fundamental physical mechanisms of thermostability.

Convergent evolution represents a phenomenon when two distinct species with differing ancestries evolve to display similar physical features (13). Environmental circumstances that require similar developmental or structural alterations for the purposes of adaptation can lead to convergent evolution even though the species have different origin. As a consequence of convergent evolution, biological structures or species that exhibit similar functions or/and appearance may appear, even though they evolved through widely divergent evolutionary pathways and had different ancestors. These similarities are typically explained as the result of common adaptive solutions to similar environmental pressures on the level of the organism. On the protein level, these adaptation similarities that arise as a result of the same selective pressures and unfortunately can be misleading to understanding the natural evolution. Therefore, identification of specific residues or fragments which may be more relevant to protein thermostability is influenced by the possibility that some of the differences among the thermophiles and mesophiles rely on phylogenetic differences instead of thermal adaptation or *vice versa* (14).

Thermostability properties of proteins gave origin to many studies so far. However, up to date no general physical mechanism was found that can be named as the most important factor for increased thermostability. Hyperthermophiles belong to two phylogenetically very different domains of life, the Bacteria and Archaea. Therefore, the strategies of molecular mechanisms including heat adaptation may be rather dissimilar depending on the phylogenetic position of the corresponding organism.

1.4.Clusters of Orhologous Groups

Proteins with elevated stability properties are not just present in organisms that survive in extreme temperature conditions, but in other biological systems from all domains of life, being involved in vast range of cellular processes where thermophilic character is, or was - evolutionary speaking, at some point essential. Discussing biological function of identified proteins with highly stable character from different unrelated organisms is possible. Therefore, The Clusters of Orthologous Groups of

proteins (COGs) database has been designed as an attempt to classify proteins from completely sequenced genomes on the basis of the orthology concept (15). Many proteins are members of paralogous gene families and have significant matches in other species. The genes in all genomes are derived from a set of unique ancestral genes present in a progenitor of all organisms. The COGs database relies on phylogenetic classification of the all the proteins encoded in complete sequenced genomes of Bacteria, Archaea and Eukarya, available at the web (http://www.ncbi.nlm.nih.gov/COG). The COG were constructed by applying the criterion of consistency of best hits specific to particular genome to the results of an exhaustive comparison of all protein sequences from those genomes (16). But, as the level of divergence between orthologous genes approaches the level of divergence among paralogs within a species, it is difficult to determine the relation between similar genes in different species. In most of the cases orthologous proteins have the same domain architecture and the same function, but there are also significant exceptions particularly among multicellular eukaryotic organisms.

COG is a functional classification based only on standard sequence-similarity method, using all-against-all sequence comparison of proteins in complete genomes. This way, comparison elucidates groups that contain a set of individual orthologous proteins or orthologous sets of paralogs from different phylogenetic lineages. Normally, orthologs are functionally equivalent proteins that arise from vertical evolution, whereas paralogs are the result of duplication events and their function may have diverted from the original ancestor. Upon duplication of an ancestral gene, copies of the gene may be subsequently lost through natural selection or simply by a neutral stochastic process. This process of duplication and divergence, along with the occasional transfer of genes between strains and species has given basis to the present contents of a genome (17). Each COG is represented by a protein with a characterized function or domain. Individual COGs are assigned to general functional categories, which represent major cellular processes, and in some cases, if known, to more specific pathways or systems. The COG functional categories are identified by one-letter codes (Fig. 1.4). Functional

classification of genes that is conserved across different organisms has provided new information about how these functions are maintained and modified across phylogenetic groups during evolution. However, in overpopulated COGs, the orthologous relationships between members are difficult to delineate precisely. Such COGs might contain proteins that evolved new functions with respect to the original ancestor, and even though these proteins still have significant sequence similarity, at the entire sequence or the domain level, they may be part of different cellular processes. Therefore, proteins involved in biological processes characteristic of eukaryotes may not have the counterparts in bacterial and archaeal genomes.

Class	COG functional classification	Class	COG functional classification
A	RNA processing and modification	N	Cell motility
B	Chromatin structure and dynamics	O	Posttranslational modification, protein turnover, chaperones
C	Energy production and conversion	P	Inorganic ion transport and metabolism
D	Cell cycle control, mitosis and meiosis	Q	Secondary metabolites biosynthesis, transport and catabolism
E	Amino acid transport and metabolism	R	General function prediction only
F	Nucleotide transport and metabolism	S	Function unknown
G	Carbohydrate transport and metabolism	T	Signal transduction mechanisms
H	Coenzyme transport and metabolism	U	Intracellular trafficking and secretion
I	Lipid transport and metabolism	V	Defense mechanisms
J	Translation	W	Extracellular structures
K	Transcription	Z	Cytoskeleton
L	Replication, recombination and repair	-	Not in COGs
M	Cell wall/membrane biogenesis		

Fig.1.4. COG functional classification categories

1.5.Designability and evolvability of protein structure

The information coded in the amino acid sequence of a protein completely determines its folded structure (18), but designability is a property measured by the number of thermodynamically foldable sequences that fold into a certain structure (19). Proteins are highly different in terms of their designability. Highly designable structures have an increased number of associated sequences that is often much larger than the average. They are also thermodynamically more stable than other structures suggesting that protein structures selected in nature are

designed and robust against mutations, and that such a selection simultaneously leads to thermodynamic stability (19, 20).

Protein structures are classified into different folds. Proteins that have the same fold also have the same major secondary structures in the same arrangement with the same topological connections, with some variations typically in the loop region. Those that are evolutionary closely related often have high sequence similarity and share a common fold, but the common fold is possible even for proteins with distinct evolutionary origins and different biological functions. Therefore, the number of folds is much lower than the number of proteins (20). The usage of protein folds in nature is known to be non-uniform: a few folds are used often, while most others are used relatively rarely.

Mutations and evolutionary selection are able to create new or improved phenotypes aiming to improving performance of various biochemical tasks, which is called evolvability of protein structure. Proteins tend to be only marginally more stable than it is required by their environment. Evolution selects for a protein's biochemical function rather than its stability. However, when protein's function depends on its ability to fold to a thermodynamically stable native structure (18, 21), stability is still a determining factor during evolution. If a protein folds at all, it must achieve its native structure with some minimal stability to remain folded at physiological conditions or by performing its physiological function. On a road to achieving functional activity, different scenarios might occur: If a protein fails to meet its minimal stability requirement, then it will fail to function. If a protein does fold with at least the minimal required stability, then evolution selects for a protein's function and is indifferent to the amount of extra stability it possesses (22). Most proteins, however, will still be marginally stable. Marginal stability relies on the fact that natural selection does not directly favor extra stability over destabilizing mutations, in the cases when protein function is preserved. Therefore, possible mechanism by which natural evolution increases evolvability is to stabilize proteins undergoing adaptive evolution or provide systems to buffer the effects of destabilizing protein mutations (22). Extra stability is not crucial in respect to selection for protein function, but it can be crucial in allowing a protein to tolerate mutations

within useful phenotypes. Most of the substitutions destabilize the native structure of a protein, therefore modest raise in thermodynamic stability increases the number and type of substitutions that a protein can tolerate before misfolding (23). Necessity for increased stability in highly expressed proteins would restrict the set of evolutionarily viable sequences and as a consequence slow sequence evolution.

Knowledge of biophysical causes of rate differences in comparing evolvability of various proteins is still scarce. A dominant factor in reducing the evolvability rate is high protein expression level that leads to increased transcription and translation (24). This increase also increases the probability of translational missense errors that may have misfolding as a consequence, and further loss of biological function. Therefore, slow evolvability is a property of highly expressed proteins in nature (25). Chaperoning cellular systems that assist in folding of other proteins and buffer various effects of mutations, therefore are also found to influence evolvability of protein structures (26) by enabling destabilizing mutations to accumulate.

1.6.Cellular environment

Various factors present in the intracellular environment influence protein folding as well as stability of mature protein. As we are dealing with properties of the proteins with highly stable behavior, introduction to an intracellular environment and it's (de)stabilizing properties gives a better insight into protein "living conditions". We should keep in mind that the subject of this thesis and research are proteins with elevated stability properties, investigated out of their natural cellular environment, but we can only assume that properties that lead to increased stability within the cell are the same ones that were proven to increase the stability in laboratory controlled environment.

Molecular crowding influence on proteins´ folding and stability

The cytoplasm of cells is an aqueous medium with a high concentration of small molecules, macromolecules and supramolecular assemblies with concentration of ~50-400 mg/ml (27, 28) where the concentrations of single species is not high. Macromolecules present in the cell occupy a

significant part of the total volume of about 40% of the medium. Therefore, the accessible volume in the cell is reduced and a significant fraction of the water is involved in solvation and does not behave as bulk water. This medium is referred to as a solution with molecular crowding. The structure and dynamics of macromolecules and supramolecular assemblies is the result of a large number of small forces such as electrostatic interactions and the hydrophobic effect. Both forces are strongly dependent on the properties of water. In this medium, a significant fraction of the water is involved in solvation and does not behave as bulk water. Therefore, various types of intermolecular forces are strongly affected by the reduced availability of water due to molecular crowding.

Crowding has a complex effect on the rate of biochemical reactions. On one hand, under crowding conditions the thermodynamic activities of the reactants increase and, on the other hand, crowding reduces diffusion and the possibility of the meeting of two reactants. The overall result of these opposing factors depends on the specific nature of each reaction. The overall decrease in the diffusional mobility of the macromolecules in the medium characterized by high viscosity should lead to a decrease in the rates of biochemical reactions in which several macromolecules are involved (29). Crowding affects all biochemical processes where changes in excluded volume are observed including processes as protein unfolding induced by chemical or thermal influence, the collapse of newly synthesized polypeptide chains into compact functional proteins, the formation of oligomeric structures and multienzymatic complex systems in metabolic pathways, protein and nucleic acid folding with formation of the compact structures, the aggregation of proteins into nonfunctional aggregates leading to some known human diseases (e.g. Parkinson's and Alzheimer's). Crowding prevents the self-assembly of partly folded polypeptide chains, stimulating the interaction between exposed hydrophobic surfaces of different chains, thus increasing the propensity of these surfaces to bind to one another (30). In biochemical reactions crowding affects reaction equilibrium by preferentially destabilizing either reactants or products - the most favored state excludes the least volume to the other

macromolecular species present in solution. Association reactions are therefore highly favored under crowded conditions, and association constants under crowded conditions could be several orders of magnitude larger than those measured in dilute solutions (31). This implies that aggregation of refolding protein molecules is a much greater problem under crowded cellular conditions than it is in dilute solutions (32).

Thermophilic and hyperthermophilic organisms generally accumulate compatible solutes as a mechanism of osmotic adjustment and protection of cell components against thermal denaturation. Newly discovered solutes from thermophilic and hyperthermophilic organisms include cyclic-2,3-bisphosphoglycerate two isomers of di-*myo*-inositol phosphate, mannosylglycerate and mannosylglyceramide di-mannosyl-di-*myo*-inositol phosphate, diglycerol phosphate and galactosyl-5-hydroxylysine (33) and may constitute an adaptive feature of these organisms to high temperatures. Thermophiles and hyperthermophiles accumulate compatible solutes that have not been found, or are rarely encountered in mesophilic organisms. Therefore, the compatible solutes of (hyper)thermophiles are specifically associated with life at high temperatures. Archaeal compatible solutes are generally negatively charged, while other microorganisms generally accumulate neutral or zwitterionic compatible solutes. Nature of interactions between solutes and exposed groups in the protein structure and its stabilizing effect is attributed mainly to a large contribution from interactions with exposed backbone groups in a partially unfolded state, with side-chain interactions modulating the specificity of the effect. The interactions should cause a contraction of the protein structure with a concomitant decrease in internal mobility (34) which is in agreement with correlation of higher thermal stability of hyperthermophilic proteins structure rigidification upon in vitro addition of compatible solute (35).

Molecular chaperones

The existence of crowding in living cells is a possible reason for the presence of chaperones for the correct folding of nascent polypeptide chains. Even though “the correct folding of polypeptide chains into the functional protein structure depends only on their amino acid sequence”

(18), the presence of molecular chaperones was found to be essential (27). In a living cell self-folding is inefficient and slow, with danger of misfolding and aggregation due to crowding conditions.

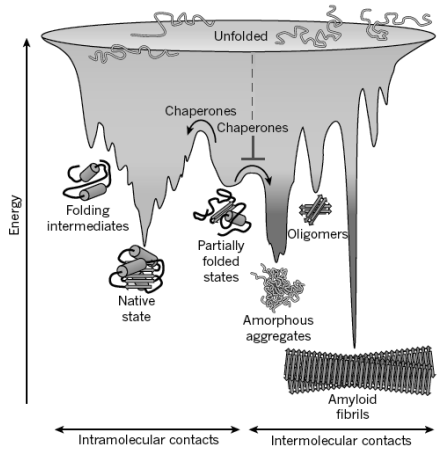


Fig.1.5. Competing reactions of protein folding and aggregation**.** Scheme of the funnel-shaped free-energy surface that proteins explore as they move towards the native state by forming intramolecular contacts. The ruggedness of the free-energy landscape results in the accumulation of kinetically trapped conformations that need to traverse free-energy barriers to reach a favorable downhill path. *In vivo*, these steps may be accelerated by chaperones. When several molecules fold simultaneously in the same compartment, the free-energy surface of folding may overlap with that of intermolecular aggregation, resulting in the formation of amorphous aggregates, toxic oligomers or ordered amyloid fibrils. Adapted from Hartl et al (30).

Molecular chaperones interact with unfolded or partially folded protein subunits, e.g. nascent chains emerging from the ribosome, or extended chains being translocated across subcellular membranes, stabilize non-native conformation and facilitate correct folding of protein subunits without interaction with native proteins or becoming part of the final folded structures (Fig. 1.5). Some chaperones are non-specific, and interact with a wide variety of polypeptide chains, but others are restricted to specific targets, they often couple ATP binding/hydrolysis to the folding process. Essential for viability, their expression is often increased by cellular stress - the heat shock response and thermal

acclimation are ubiquitous mechanisms in the extremely thermophilic microorganisms. Molecular chaperones can be classified into three functional groups based on their action mechanism: i) Folding modulators are chaperones that assist and mediate folding processes (DnaK and GroEL) performing their function on conformational changes in the presence of ATP; ii) Holding chaperones that stabilize partially folded protein structure in a case of severe stress situation (Hsp33, Hsp31 and IbpB) awaiting folding chaperones to become available; iii) Chaperone (ClpB) that promotes the solubilization of aggregated proteins as a result of stress (36).

The chaperones that are shared by Archaea and Bacteria include the chaperone machine composed of Hsp70(DnaK), Hsp40(DnaJ), and GrpE (37). In Archaea, proteins coded by these genes are very similar to bacterial homologs, as if the genes had been received via lateral transfer from Bacteria. The chaperonin system in Archaea studied to the present, including those that possess a bacterial-like chaperone machine, is similar to that of the eukaryotic-cell cytosol. Hyperthermophilic Archaea like *Pyrococcus spp, Sulfolobus spp, Pyrobaculum aerophilum, Methanocaldococcus jannaschii, Metahopyrus kandleri, Archeoglobus fulgidus* do not have Hsp90, DnaK, DnaJ, GrpE, Hsp33 and Hsp10 homologs (38).

1.7.Diversity of Protein Conformational states

Protein stability is defined by imprecise cancellation of two large effects, namely, the hydrophobic effect, favoring folding, and chain entropy, disfavoring it. Because these two effects are similar in magnitude the net stability of proteins is marginal. Indeed, the average Gibbs free energies of denaturation (ΔG_D) ranges from 20 to 60 kJ/mol (at 25^0C), a value which is comparable to the magnitude of the weak forces that stabilize the native conformation (discussed below). The fact that protein stability relies on a minor free energy difference between the native and unfolded states implies a certain conformational flexibility of the structure. Indeed, due to sufficiently low kinetic barriers, other alternative conformations, with considerably low energy, are accessible to the protein (39). One such example is the molten globule state, which

is almost as compact as the native form and have a loosely packed core, while retaining some of its native secondary structure (40). Consequently, each protein has a particular energetic landscape of conformations that it can adopt under physiological conditions. Ultimately this may correspond to a protein in which the most populated ensemble of structural conformations is disordered or that contains highly disordered regions (39, 41). This possibility impacts on the structure-function paradigm, as disordered proteins are biologically active in functions related to regulation of transcription and translation, protein phosphorylation, storage of small molecules, and regulation of the self assembly of multiprotein complexes (42). Disordered proteins often fold into an ordered structure upon binding to a protein partner. That is, for example, the case of thyroid hormone and retinoid receptors, which occur as unstructured ensembles, and fold upon interacting with a nuclear receptor binding domain. The same occurs upon interaction with a ligand, such as a metal ion or a nucleic acid, as in the zinc-finger containing transcription factor TFIIIA or the translation initiation factor elF4E (42, 43). Interestingly, disordered domains may adopt distinct ordered conformations depending on the interacting partner, thus reflecting a significant functional flexibility, in agreement with the fact that disordered proteins are mainly involved in signaling and regulatory pathways. This has led to the proposal of generically coining disordered proteins as 'pliable' (43). Overall, some proteins comprising intrinsically disordered domains or segments have a plethora of accessible ordered conformations that they can adopt, upon protein-ligand interaction(s). In thermodynamic terms, coupling folding to binding results in a strategy to use the binding enthalpy to pay the entropic cost to fold a disordered protein.

1.8.Molecular determinants of protein stability

The dominant forces that fold proteins and stabilize the native states have long been identified (44). These comprise residue-residue interactions and water-residue interactions, and are typically classified into the following classes: van der Waals interactions (present between any group), hydrogen bonds, salt bridges (bonds between oppositely charged residues that are sufficiently close to each other to experience

electrostatic attraction), disulfide and hydrophobic forces (45). The magnitudes of these forces in proteins, as determined from mutagenesis and unfolding experiments, are very low. For example, a stabilising contribution of 4-5 kJ/mol has been estimated per hydrogen bond (46); a single ion pair may be responsible for a 12.5-20 kJ/mol stabilization (47). Short distance (<7Å) aromatic pairs interactions, such as Tyr-Tyr and Phe-Phe, contribute approximately 5.5 kJ/mol towards thermodynamic stabilization (48). However, the fact that the free energy of folding is always very low (20 to 60 kJ/mol), makes these low-magnitude forces relevant contributors to the overall stability, as the presence or absence of one, or a few, of these interactions may be enough to shift the equilibrium towards the native or unfolded state. Among these, hydrophobic interactions are widely believed to be the main driving forces behind the folding of globular proteins (49) and result from the incapacity of non-polar residues to make hydrogen bonds with water molecules. The resulting water-residue interactions are thermodynamically unfavorable and drive the non-polar residues into the interior of proteins (50, 51) where they arrange into densely packed clusters. The contribution of the hydrophobic effect to the stability of globular proteins has been estimated from cavity-creating mutations in which a small aliphatic chain in the interior of the protein replaced a larger one with identical characteristics: this has shown that an energetic gain of ~5.5 kJ/mol is obtained per buried methyl group (46). Many proteins are also stabilized by covalent interactions which crosslink segments of secondary structure in the native state: these comprise disulfide bridges, coordination of a metal ion or attachment of an organic or organometallic cofactor (52). Secreted and periplasmatic proteins, or those that are present at external surfaces such as cell walls, have an increased content of disulfide bridges. These confer an increased conformational rigidity to the protein in the oxidizing conditions found outside the cell. Metal ion cross linked domains are found in zinc fingers (53), which occur in transcription factors and constitute the most abundant domain encoded in the human genome, or in iron-sulfur proteins, namely ferredoxins. In both examples, removal of these ions frequently results in protein unfolding or in the formation of non-native states (54, 55). While the major contributors to protein stability are non-

covalent interactions, metal ion and disulfide cross-linking interactions assume a particularly relevant role in small proteins comprising irregular domains. Although proteins frequently lack a large hydrophobic core and have minimal secondary structure; covalent interactions allow attaching and stabilizing different parts of the protein. This is what happens, e.g., with scorpion toxin, allergy factor Ra5, and several protease inhibitors (52).

1.9.Proteins with enhanced conformational stability

Some proteins need to be particularly stable as a result of molecular adaptation to a particular physiological condition or to a harsh environmental factor, such as high salinity, extreme pH and high temperatures. Among thermophiles, enhanced protein stability encompasses both thermodynamic and kinetic stability. While the kinetic stability depends on the energy barrier to unfolding, i.e., on the activation energy of unfolding, the thermodynamic stability is reflected in the conformational stabilities (i.e., ΔG_D), which may be up to 100 kJ/mol larger than those from mesophilic proteins (56), and in the midpoint transition temperatures for unfolding (i.e., T_m), which are typically between 20-30^{0}C above those of mesophiles (57, 58). A recent study suggested that the thermodynamic strategy leading to the higher denaturation temperatures exhibited by thermophiles relies on the elevation of the stability curve (i.e., ΔG vs. T) (58), rather than on broadening, or shifting it toward higher temperatures. A relevant problem in protein chemistry is that of identifying and understanding the intrinsic factors that are responsible for the functional stability exhibited by thermophilic proteins at very high temperatures. This fundamental problem of paramount importance was initially investigated by Perutz and others in the 1970s (50), and since then, numerous studies have tackled this issue. The following structure-based approaches can be loosely identified: (i) comparison of structures from thermophilic proteins with those of their mesophilic homologues and (ii) systematic structure-based sequence comparisons for a group of proteins. (iii) More recently, due to progress in genome sequence projects, it has become possible to perform large-scale comparison between genome sequences from thermophiles and mesophiles. Based on experimental and

theoretical analyses, diverse stabilizing strategies have been suggested, as putative intrinsic drivers for the enhanced stability exhibited by thermophilic proteins (7). Structural properties like better core (and secondary structure) packing (59-61), deletion or shortening of loops and increased helical content have been typically ascribed to thermophilic proteins (62). On the physicochemical side, it is generally claimed that thermophilic proteins have more hydrophobic residues (63-66), a larger amount of main-chain hydrogen bonds (66) and a higher number of proline residues. However, a recent study, which analyzed 18 non-redundant families of thermophilic and mesophilic proteins, reported that these factors do not show consistent, substantial variations between mesophiles and thermophiles (67). The higher number of salt bridges among thermophilic proteins suggests that these interactions play a role in stability enhancement (62, 66-71). Nevertheless, there is not a general physical mechanism able to rationalize the stability enhancement upon increasing the number of salt bridges, because the net electrostatic free energy of salt bridges can be either stabilizing or destabilizing (72). This results from the fact that energetically favorable Coulombic charge–charge interaction forming in the protein core, is opposed by the unfavorable desolvation of interacting charges – the transfer of a salt bridge from water to nonpolar environment costs ~ 42-67 kJ/mol (73). Therefore, it is mostly the surface, solvent-exposed, salt bridges that effectively lead to an increase in protein stabilization (74-77). Moreover, there is also some evidence that extended networks of salt-bridges (formed by residues that participate in more than one salt-bridge) between protein subunits are critical for achieving the superior thermostability of hyperthermostable proteins (63, 67, 78).

1.10. Factors influencing protein stability

In order to investigate origins of elevated protein stability in the soluble cellular proteome of chosen organisms in this study, presence/absence of various stabilizing factors have been taken into account, among them intrinsic and structural factors. Therefore, information previously published in the literature is here briefly presented as the insight into stabilizing strategies from biological point of view. It is important to remind that none of these factors have been explored in this thesis with

an attempt to find new prediction method of theoretical analysis for thermostable proteins.

1.10.1. Intrinsic factors influencing protein stability

Amino acid composition

Amino acid composition has long being known to have a certain influence on proteins' thermostability. Comparing thermophiles and mesophiles, difference of the amino acid composition has been found to be a global trend across a large number of protein families (67). Statistical analyses comparing amino acid compositions in mesophilic and thermophilic proteins indicated trends toward substitutions of certain amino acids with the others that implied in order to increase thermostability. Some of it has been indicated by various studies although the reasons behind the origins of these substitutions are still open for discussion. Many factors might substantially influence the results as well as the conclusions: chosen sample group of thermophilic vs. mesophilic organisms, selection of their representative proteins or availability of data for the comparison. However, the majority of studies would agree to some overall conclusions regarding the frequency of certain amino acids.

Charged residues such as Glu, Asp, Lys, and Arg are frequent and polar residues such as Ser, Thr, Asn, and Gln, are scarce in thermophilic proteins (79, 80). Polar residues Ser, Thr and Asn can hydrogen bond to the backbone peptide groups and this interferes with the hydrogen bonding of the α-helix; therefore these amino acids are found less frequently in α-helices, and at the same time less often in thermostable proteins. Substitutions such as Gly→Ala and Lys→Arg are frequent in thermophiles, where higher alanine content in thermophilic proteins is explained by the fact that Ala was the best helix-forming residue (7). Hyperthermophilic proteins also contain slightly more hydrophobic and aromatic residues than mesophilic proteins do. Frequency of the thermolabile residue Cys and of Ser was found to be low in thermophilic proteins. Cys is a hydrophobic amino acid with a short side chain containing one sulfur atom, which is believed to be unfavorable for thermostability as sulfur atom is relatively large and cannot contribute to

the compact packing of protein structures. On the whole, a higher proportion of amino acids in the thermophilic proteins adopt α-helical conformation (67). These data obtained from genome sequencing cannot be generalized, since large variations exist among hyperthermophile genomes themselves. Applying temperature beyond 100 °C the thermal stabilities of the common amino acids are (Val,Leu) >Ile>Tyr>Lys>His>Met>Thr>Ser>Trp>(Asp,Glu,Arg,Cys) (81). As the same amino acids serve as building blocks for both mesophilic and (hyper)thermophilic proteins the interaction between the amino acids in the peptide chain is a key feature for increased stability. Protein structures can be stabilized by decrease in their entropy of unfolding (82). Proline occur less frequently in α-helices of the thermophilic proteins as it can adopt only a few configurations and can restrict the configurations allowed for the preceding residue, so it has the lowest conformational entropy. In the unfolded state, glycine, without a β-carbon, is the residue with the highest conformational entropy (7). Mutations Gly→Xaa or Xaa→Pro should decrease the entropy of a protein's unfolded state and stabilize the protein, as long as substituted residue does not introduce unfavorable strains in the protein structure. Replacement of other residues by Pro at suitable positions may enhance protein thermostability (83) in thermophilic proteins, especially in the loops. Frequency of the aromatic amino acids such as Phe, Trp, and Tyr in both thermophilic and mesophilic proteins were investigated in various studies that have reported having higher frequency of Tyr in thermophiles compared with mesophilic proteins (67) and in some cases elevated number of aromatic pairs (7). But, as for the other types of interactions, localization rather than frequency of certain residues and clusters has often been found to influence protein stability. The most of the additional aromatic clusters in the thermophiles that improve thermostability are not part of the protein core and they occur as separate entities on the protein surface. In these cases, aromatic residues forming the additional cluster stabilize the structure by the tertiary fold. Also, the presence of additional aromatic clusters near the active site help in retaining the conformational features of the active sites that is required to bind the substrate at high temperatures and thus contributing to the high thermophilicity of the thermostable proteins.

Additional aromatic clusters occur in regular secondary structures, implying their location to be in more rigid regions of the protein (84).

Besides from these studies that are focused solely on amino acid composition, some results indicate that frequency of certain combination of amino acids influences thermal stability. Prevalence of either residue combination (IVYWREL)(85) or their ratio ((E+K)/(Q+H))(86) has been noted to lead to thermal stability. It is very important to note on the level of sequence-encoded increment of thermostability that not just the frequency but the position of certain residues has significant influence.

Intracellular and extracellular proteins have different amino acid compositions as expected, and therefore, their subcellular locations of the protein subpopulation investigated, highly influences the result (87). These differences are consequence almost entirely to residues exposed to the solvent (88). Subcellular location determines the differences in the amino acid composition resulting from the adaptation of the protein surface to the physicochemical environment. Therefore it would be reasonable to expect that the differences in the amino acid compositions between proteins of thermophilic and mesophilic organisms would be much greater on the protein surface than the interior, regarding the differences between the environments in which thermophilic and mesophilic organisms live (79). Therefore, differences in amino acid composition between organisms with different optimal growth temperature (OGT) might often be evolutionarily relevant, rather than an indication of its adaptation to high temperatures - more relevant to thermostability than amino acid composition is the distribution of the residues and their interactions in the protein. With increasing experimental data accumulating, in particular complete genome sequences, it is already obvious that thermophilic adaptation cannot just be defined in relation to significant differences in the amino acid composition (7).

Ion pairs

Electrostatic interactions are an important factor influencing thermostability of proteins (68), This opinion is supported by the

increased number of salt bridges found in many structures of thermostable and, in particular, hyperthermostable proteins. These results are mostly based on structural comparisons of proteins from hyperthermophiles and mesophiles and have suggested that the ion pairs show a tendency to be increased in number as well as to be organized in large networks, in parallel with increasing melting temperature of the protein (50). As previously mentioned, comparative analysis using complete genome sequences from Bacteria, Eukarya and Archaea, a large difference between the proportions of charged amino acids such as Arg, Asp, His, Glu, and Lys versus polar amino acids was found to be a specific signature of all proteins from hyperthermophilic organisms (60, 80). Proteins from hyperthermophilic organisms are characterized by an increased number of ion pairs with respect to the statistical expectance and/or the number of ion-pairs in their mesophilic counterparts. This suggests that electrostatic interactions emerge as a factor responsible for the elevation of the melting temperature of proteins from hyperthermophilic organisms. It is important to note that, in addition to the increased number of charged group, their spatial organization influences elevation of the melting temperature of thermostable proteins. In parallel with increasing Tm of the protein, ion pairs tend to be organized in networks that are often found on the protein surface or partially buried at domain or subunit interfaces near to local symmetry axes of protein oligomers. The largest ion-pair network in any hyperthermostable protein reported has been observed in glutamate dehydrogenase from *P. furiosus* (63, 67). Statistical appearance of salt bridges is higher for proteins without disulfide bridges than for proteins with disulfide bridges which is explained by the fact that salt bridges often occur between groups distant in the protein sequence, forming cross-links that therefore stabilize the tertiary structure (80). By contrast, charge burial in the protein environment is energetically unfavorable and can reduce the energy gained by attraction between opposite charges in an ion pair.

In general, the optimization of electrostatic interactions is gained mainly by minimizing repulsive contacts, not by increasing the salt bridge number. The proteins from hyperthermophilic organisms optimize their

electrostatic interactions in both directions: by minimizing repulsive contacts and by increasing the number of ion pairs, so that they gain electrostatic stabilization by minimizing the number of repulsive contacts rather than by creating salt bridges (68).

Hydrophobic interactions

Hydrophobic effect is the main driving force in protein folding (46). Aliphatic amino acids such as Ala, Ile, Leu, and Val in both thermophilic and mesophilic proteins contribute to the hydrophobic interaction, which is main force for maintaining conformational stability in inner part of protein (60). It has been suggested that thermophilic proteins are substantially more hydrophobic and have more surface area buried upon oligomerization as compared with their mesophilic counterparts, even though overall hydrophobicity of thermophilic proteins and their mesophilic homologs are very similar (67). Therefore, it is important to compare the frequency of the residues with the same solvent accessibility. In thermophilic proteins, the amino acid with the short alkyl group tend to interact more closely with neighbouring residues and have better packed form in protein structure. Ala and Val that are surface exposed with lower frequency but high frequent in well-buried state contribute more to protein thermostability by enhancing conformational stability in the buried part of protein structure (60). With packing, the hydrophobic effect can have consequences at the level of the individual protein chains due to larger and more hydrophobic protein core but also due to the association of the chains (7). Overall numbers of water molecules per protein molecule and the water-accessible surface area is larger in the mesophilic species proteins than in their thermophilic homologs (89).

Disulfide bridges

Frequency of Cys residues in thermophilic proteins is significantly decreased due to two reasons. Thermostability has, among other parameters, been attributed to enhanced secondary structure propensity. Cys is a helix disfavoring residue, and analysis of the composition of α-helices in the thermophilic proteins (90) have noted its

significant decrease, together with His. The second reason is that Cys, together with Asn, Gln and Met can be classified as thermolabile due to their tendency to undergo deamidation or oxidation at high temperatures (67). Even though disulfide bonds are not a method to achieve protein thermostability (80) and in general, hyperthermostable proteins contain lower fractions of cysteines and less disulfide bonds than their thermostable and mesostable counterparts, there still are some rare evidence of stabilization due to disulfide bond like in Fd from hyperthermophile *Arquiflex aeolicus* (91), although not as the only mechanism of stabilization but rather in contribution to electrostatic interactions by enabling dimerization of the subunits (92). Even though the other examples of thermal stabilization of proteins by introduction of disulfide bonds exist, it is obvious that is not the strategy chosen by thermophilic but rather by mesophilic organisms. Fewer cysteines are present in thermophilic proteins due to their property of being the most reactive amino acids in proteins. Their autooxidation, usually catalyzed by metal cations, especially copper, leads to the formation of intramolecular and intermolecular disulfide bridges or to the formation of sulfenic acid. Disulfide bond reshuffling can cause important structural variations. Therefore, after forming incorrect intersubunit disulfide bridges protein often becomes less stable and less thermophilic than the native enzyme (93).

Hydrogen bonds

Significant difference in the number of hydrogen bonds among mesophilic, thermophilic, and hyperthermophilic enzymes but the dependence lies mostly in their position within the structure, in order to influence thermostability. According to this, there are two reasons why this type of H bond might be particularly thermodynamically stabilizing: (i) the desolvation penalty associated with burying such H bonds is less than the desolvation penalty for burying an ion pair (that involves two charged residues), and (ii) the enthalpic reward of a charged-neutral H bond is greater than that of a neutral-neutral H bond because of the charge-dipole interaction (94). This correlation between charged-neutral H bonds and protein stability suggests that the role of charged residues in protein stabilization may not be limited to forming ion pairs. An increased

number of charged-neutral H bonds was also found in the *T. maritima* ferredoxin that either stabilize the structure of turns or anchor turns to one another (7). On the other hand, hydrogen bonding that was compared among the nine glycosidases originating from *Thermus nonproteolyticus* by dividing into three classes: main chain-main chain (MM H-bonds), main chain-side chain (MS H-bonds), and side chain-side chain (SS H-bonds) (75) gave other conclusions. Therefore, the position within the protein and not the difference in the number of hydrogen bonds influences the most thermal stability properties.

1.10.2. Structural and other extrinsic factors influencing protein stability

Protein classes

Depending on the chosen protein samples, available amount of data and strategy of their analysis, it has been found in the literature that some protein classes (α, β, $\alpha+\beta$, α/β) are more frequent among proteins with elevated stability properties, while some are less common.

Previous studies have found increase in helix content is greater in moderately thermophilic proteins whereas the increase in β content is much more significant in extremely thermophilic ones, while both groups are strongly dominated by α/β type proteins (71). Protein stability can be determined by the stability and tight packing of its core, therefore the propensity of the individual residues to participate in helical or strand structures can be taken as a potential stability mechanism. Helix-favouring residue Arg occurs more frequently in α-helices of thermophilic proteins, whereas helix-disfavoring residues Cys and His have lower frequencies of occurrence in thermophilic helices (67). The appearance of an increased number of charged-neutral H bonds was found in the T. maritima ferredoxin, that was explained in away that these H bonds either stabilize the structure of turns or anchor turns to one another (7). The frequency of the general secondary structural composition and the secondary structural location of the aromatic residues that form the additional aromatic clusters, happen to be located in more rigid regions of the protein (84). According to this study almost 38% of the aromatic

residues in the additional aromatic clusters form helices, 32% strands, 21% coil and 9% from the loop regions of the thermophilic proteins. However, the secondary structural composition of the thermophilic proteins has nearly 54% in coils or loops and nearly 26% are strands and only 20% are helices. Various types of conformational strain releases have been proposed as stabilizing mechanisms. In α-helices, for example, residues with a low helical propensity can be replaced by residues that have a high helical propensity (7). Such substitutions usually take place when a residue's side chain is not well accommodated in the α-helix. One particular substitution location in α-helices is the C terminus. Gly is the most favorable residue there, because its lack of side chain allows it to adopt a left-handed helical conformation without strain and because the main chain carbonyl oxygen can form H bonds with solvent molecules. The *P. furiosus* citrate synthase contains at least seven helices that have Gly at C terminus (95). In general, these types of conformational strain releases are not expected to provide significant stabilization, and they have not been characterized in detail in hyperthermophilic protein structures. They also compete with other stabilization mechanisms such as propensity for hydrophobic interactions, for H bonds, or for ion pairs. To conclude, it can only be noted beyond reasonable doubt that irregular structural regions are much less frequent in highly stable structures, while any class (α, β, $\alpha+\beta$, α/β) has its advantages and disadvantages, and at least for now can not be taken as the definitive indication of elevated stability.

Reduction of solvent accessible surface, oligomeric state

The packing density is defined as the ratio of the volume enclosed by the van der Waals envelope of a given atom to the actual volume it occupies. It is often increased in hyperthermophilic proteins leading to a lower thermal mobility in these proteins compared to the mesophilic proteins at a given temperature (96). The increased rigidity of hyper-thermophilic proteins at mesophilic temperatures may be result of an increased compactness, which can be obtained by a decrease in both number and size of internal cavities. Surface to volume ratio has been shown to be significantly lowered thermophilic vs. mesophilic proteins, a decrease in

the solvent accessible area has been observed in proteins with an increased thermostability (63).

An increasing number of hyperthermophilic proteins are known to have a higher oligomerization state than their mesophilic homologues (7). Among oligomeric proteins, including homomers and heteromers, the most common assemblies are homodimers (97). In the study performed with mesophilic proteins, based on stability studies of dimeric globular proteins, it was calculated that quaternary interactions could provide 25 to 100% of the conformational stability in protein dimmers (98) and suggested that oligomerization can be a significant stabilizing mechanism for hyperthermophilic enzymes. The large stabilization energy of dimers is primarily due to the intersubunit interactions. The magnitude of the conformational stability is related to the size of the polypeptide in the subunit and depends upon the type of structure in the subunit interface. Protein–protein interfaces are characterized by their biochemical properties: % of hydrophobicity, polarity, hydrogen bonds, and geometrical properties: interface size, shape, atomic packing, planarity and complementarity (99). Hydrophobic interactions stabilize the interface, ionic interactions and hydrogen-bonding influence the selectivity of the interface and charged pairs may also protect against promiscuous homodimeric interactions (100). The relatively large interface areas of proteins that form a dimeric intermediate may be particularly difficult to dissociate.

Extrinsic stabilization of proteins may be achieved by influence of one or various factors such as metal ions, ligands, solvent, substrate(s), posttranslational modification and compatible solutes (101). Extrinsic stabilizing factors have not been object of this thesis, but as their influence on protein stability is documented in the literature, they deserve to be mentioned here, even though they will not be discussed in detail.

Substrate molecules have long been known to stabilize enzymes specifically by stabilizing their active site, that is also valid for some hyperthermophilic enzymes (102). Metal ions are present in vast number of presently known proteins and in many cases are usual stabilizers of

their folded structure and/or physiological activity like Zn, Mg and Ca, while Fe, Cu and Mn play role in many redox processes with electron transfer. Na and K are included in cellular response. Metal dependent protein folding involves polypeptide chain with one or more metal ions into a stable and functional fold. Metal ions contribute in achieving and maintenance of a specific fold by introducing conformational restrictions that decrease entropy of the unfolded state and favor folding reaction. In studies comparing stability of analog metaloproteins from meso-, thermo- and hyperthermophilic organisms it has been evidenced significant role of metal in structural stability (103, 104).

Most proteins undergo co- and/or post-translational modifications. Knowledge of these modifications is extremely important because they may alter physical and chemical properties, folding, conformation distribution, stability, activity, and consequently, function of the proteins. Moreover, the modification itself can act as an added functional group. Posttranslational modifications are ubiquitous in the cell and regulate the function of proteins often by modulating their biophysical characteristics. These modifications (e.g., glycosylation, phosphorylation, methylation) can adjust the thermodynamic, kinetic, and structural features of proteins. Examples of the biological effects of protein modifications include phosphorylation for signal transduction, ubiquitination for proteolysis, attachment of fatty acids for membrane anchoring and association, glycosylation for protein half-life, cell-cell and cell-matrix interactions, etc. Reversible protein phosphorylation on serine, threonine or tyrosine residues, is one of the most important and well-studied post-translational modifications that plays critical roles in the regulation of many cellular processes including cell cycle, growth, apoptosis and signal transduction pathways. Regarding influence on protein stability, scarce number of examples are known of hyperthermophilic proteins that are glycosylated, but eucaryal proteins showed that glycosylation could cause significant thermal stabilization without affecting the protein folding pathways or their conformations (81).

1.11. References

1. **Brock, T. D., and H. Freeze.** 1969. Thermus aquaticus gen. n. and sp. n., a nonsporulating extreme thermophile. J Bacteriol **98:**289-97.
2. **Maheshwari, R., G. Bharadwaj, and M. K. Bhat.** 2000. Thermophilic fungi: their physiology and enzymes. Microbiol Mol Biol Rev **64:**461-88.
3. **Kristjansson, J. K., Stetter,K.O.** 1992. Thermophilic bacteria. CRC Press, Boca Raton, FL.
4. **Stetter, K. O.** 1996. Hyperthermophiles in the history of life. Ciba Found Symp **202:**1-10; discussion 11-8.
5. **Stetter, K. O.** 2006. Hyperthermophiles in the history of life. Philos Trans R Soc Lond B Biol Sci **361:**1837-42; discussion 1842-3.
6. **Woese, C. R., O. Kandler, and M. L. Wheelis.** 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A **87:**4576-9.
7. **Vieille, C., and G. J. Zeikus.** 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev **65:**1-43.
8. **Kurosawa, N., Y. H. Itoh, T. Iwai, A. Sugai, I. Uda, N. Kimura, T. Horiuchi, and T. Itoh.** 1998. Sulfurisphaera ohwakuensis gen. nov., sp. nov., a novel extremely thermophilic acidophile of the order Sulfolobales. Int J Syst Bacteriol **48 Pt 2:**451-6.
9. **She, Q., R. K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M. J. Awayez, C. C. Chan-Weiher, I. G. Clausen, B. A. Curtis, A. De Moors, G. Erauso, C. Fletcher, P. M. Gordon, I. Heikamp-de Jong, A. C. Jeffries, C. J. Kozera, N. Medina, X. Peng, H. P. Thi-Ngoc, P. Redder, M. E. Schenk, C. Theriault, N. Tolstrup, R. L. Charlebois, W. F. Doolittle, M. Duguet, T. Gaasterland, R. A. Garrett, M. A. Ragan, C. W. Sensen, and J. Van der Oost.** 2001. The complete genome of the crenarchaeon Sulfolobus solfataricus P2. Proc Natl Acad Sci U S A **98:**7835-40.
10. **Matte, A., J. Sivaraman, I. Ekiel, K. Gehring, Z. Jia, and M. Cygler.** 2003. Contribution of structural genomics to understanding the biology of Escherichia coli. J Bacteriol **185:**3994-4002.
11. **Blattner, F. R., G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete

genome sequence of Escherichia coli K-12. Science **277:**1453-74.

12. **Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A. Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser.** 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature **399:**323-9.
13. **Stevenson, J. C.** 1991. Dictionary of concepts in physical antropology. Greenwood Press.
14. **Lin, Y. S.** 2008. Using a strategy based on the concept of convergent evolution to identify residue substitutions responsible for thermal adaptation. Proteins **73:**53-62.
15. **Tatusov, R. L., E. V. Koonin, and D. J. Lipman.** 1997. A genomic perspective on protein families. Science **278:**631-7.
16. **Tatusov, R. L., M. Y. Galperin, D. A. Natale, and E. V. Koonin.** 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res **28:**33-6.
17. **Zhang, J.** 2003. Evolution by gene duplication: an update. TRENDS in Ecology and Evolution **18:**292-298.
18. **Anfinsen, C. B.** 1973. Principles that govern the folding of protein chains. Science **181:**223-30.
19. **Li, H., R. Helling, C. Tang, and N. Wingreen.** 1996. Emergence of preferred structures in a simple model of protein folding. Science **273:**666-9.
20. **Helling, R., H. Li, R. Melin, J. Miller, N. Wingreen, C. Zeng, and C. Tang.** 2001. The designability of protein structures. J Mol Graph Model **19:**157-67.
21. **Dobson, C. M.** 2003. Protein folding and misfolding. Nature **426:**884-90.
22. **Bloom, J. D., S. T. Labthavikul, C. R. Otey, and F. H. Arnold.** 2006. Protein stability promotes evolvability. Proc Natl Acad Sci U S A **103:**5869-74.
23. **Bloom, J. D., J. J. Silberg, C. O. Wilke, D. A. Drummond, C. Adami, and F. H. Arnold.** 2005. Thermodynamic prediction of protein neutrality. Proc Natl Acad Sci U S A **102:**606-11.
24. **Rocha, E. P., and A. Danchin.** 2004. An analysis of determinants of amino acids substitution rates in bacterial proteins. Mol Biol Evol **21:**108-16.

25. **Drummond, D. A., J. D. Bloom, C. Adami, C. O. Wilke, and F. H. Arnold.** 2005. Why highly expressed proteins evolve slowly. Proc Natl Acad Sci U S A **102:**14338-43.
26. **Tokuriki, N., and D. S. Tawfik.** 2009. Stability effects of mutations and protein evolvability. Curr Opin Struct Biol **19:**596-604.
27. **Ellis, J.** 1987. Proteins as molecular chaperones. Nature **328:**378-9.
28. **Fulton, A. B.** 1982. How crowded is the cytoplasm? Cell **30:**345-7.
29. **Chebotareva, N. A., B. I. Kurganov, and N. B. Livanova.** 2004. Biochemical effects of molecular crowding. Biochemistry (Mosc) **69:**1239-51.
30. **Hartl, F. U., and M. Hayer-Hartl.** 2002. Molecular chaperones in the cytosol: from nascent chain to folded protein. Science **295:**1852-8.
31. **Zimmerman, S. B., and A. P. Minton.** 1993. Macromolecular crowding: biochemical, biophysical, and physiological consequences. Annu Rev Biophys Biomol Struct **22:**27-65.
32. **Van den Berg, B., R. Wain, C. M. Dobson, and R. J. Ellis.** 2000. Macromolecular crowding perturbs protein refolding kinetics: implications for folding inside the cell. Embo J **19:**3870-5.
33. **Lamosa, P., A. Burke, R. Peist, R. Huber, M. Y. Liu, G. Silva, C. Rodrigues-Pousada, J. LeGall, C. Maycock, and H. Santos.** 2000. Thermostabilization of proteins by diglycerol phosphate, a new compatible solute from the hyperthermophile Archaeoglobus fulgidus. Appl Environ Microbiol **66:**1974-9.
34. **Liu, Y., and D. W. Bolen.** 1995. The peptide backbone plays a dominant role in protein stabilization by naturally occurring osmolytes. Biochemistry **34:**12884-91.
35. **Lamosa, P., D. L. Turner, R. Ventura, C. Maycock, and H. Santos.** 2003. Protein stabilization by compatible solutes. Effect of diglycerol phosphate on the dynamics of Desulfovibrio gigas rubredoxin studied by NMR. Eur J Biochem **270:**4606-14.
36. **Baneyx, F., and M. Mujacic.** 2004. Recombinant protein folding and misfolding in Escherichia coli. Nat Biotechnol **22:**1399-408.
37. **Macario, A. J., and E. C. de Macario.** 1999. The archaeal molecular chaperone machine: peculiarities and paradoxes. Genetics **152:**1277-83.
38. **Karlin, S., L. Brocchieri, A. Campbell, M. Cyert, and J. Mrazek.** 2005. Genomic and proteomic comparisons between bacterial and archaeal genomes and related comparisons with the yeast and fly genomes. Proc Natl Acad Sci U S A **102:**7309-14.

39. **Uversky, V. N.** 2002. Natively unfolded proteins: a point where biology waits for physics. Protein Sci **11:**739-56.
40. **Ptitsyn, O. B.** 1995. Structures of folding intermediates. Curr. Opin. Struct. Biol. **5:**74-78.
41. **Dunker, A. K., M. S. Cortese, P. Romero, L. M. Iakoucheva, and V. N. Uversky.** 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. Febs J **272:**5129-48.
42. **Dyson, H. J., and P. E. Wright.** 2005. Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol **6:**197-208.
43. **Tompa, P.** 2002. Intrinsically unstructured proteins. Trends Biochem Sci **27:**527-33.
44. **Kauzmann, W.** 1959. Chemical Specificity in Biological Systems Rev. Mod. Phys. **31:**549-556.
45. **Chandler, D.** 2005. Interfaces and the driving force of hydrophobic assembly. Nature **437:**640-647.
46. **Pace, C. N., B. A. Shirley, M. McNutt, and K. Gajiwala.** 1996. Forces contributing to the conformational stability of proteins. Faseb J **10:**75-83.
47. **Anderson, D. E., W. J. Becktel, and F. W. Dahlquist.** 1990. pH-induced denaturation of proteins: a single salt bridge contributes 3-5 kcal/mol to the free energy of folding of T4 lysozyme. Biochemistry **29:**2403-8.
48. **Serrano, L., M. Bycroft, and A. R. Fersht.** 1991. Aromatic-aromatic interactions and protein stability. Investigation by double-mutant cycles. J Mol Biol **218:**465-75.
49. **Dill, K. A.** 1990. Dominant forces in protein folding. Biochemistry **29:**7133-55.
50. **Perutz, M. F., and H. Raidt.** 1975. Stereochemical basis of heat stability in bacterial ferredoxins and in haemoglobin A2. Nature **255:**256-9.
51. **Sandelin, E.** 2004. On Hydrophobicity and Conformational Specificity in Proteins. Biophysical Journal **86:**23-30.
52. **Petsko, G., and D. Ringe.** 2004. Protein Structure and Function. New Science Press.
53. **Auld, D. S.** 2001. Zinc coordination sphere in biochemical zinc sites. Biometals **14:**271-313.
54. **Leal, S. S., and C. M. Gomes.** 2007. Studies of the molten globule state of ferredoxin: Structural characterization and implications on protein folding and iron-sulfur center assembly. Proteins.
55. **Lee, M. S., J. M. Gottesfeld, and P. E. Wright.** 1991. Zinc is required for folding and binding of a single zinc finger to DNA.

FEBS Lett **279:**289-94.

56. **Jaenicke, R.** 2000. Stability and stabilization of globular proteins in solution. J. of Biotechnology **79:**193-203.
57. **Rees, D. C., Robertson A.D.** 2001. Some thermodynamic implications for the thermostability of proteins. Protein Science **10:**1187-1194.
58. **Razvi, A., Scholtz, J.M.** 2006. Lessons in stability from thermophilic proteins. Protein Sci. **15:**1569-1578.
59. **Querol, E., PerezPons J.A., MozoVillarias A.** 2006. Analysis of protein conformational characteristics related to thermostability Protein Engineering **9:**265-271.
60. **Pack, S. P., and Y. J. Yoo.** 2004. Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. J Biotechnol **111:**269-77.
61. **Gromiha, M. M., Oobatake, M., Sarai, A. .** 1999. Important amino acid properties for enhanced thermostability from mesophilic to thermophilic proteins Biophysical Chemistry **82:**51-67.
62. **Chakravarty, S., Varadarajan, R.** 2002. Elucidation of factors responsible for enhanced thermal stability of proteins: A structural genomics based study. Biochemistry **41:**8152-8161.
63. **Yip, K. S., K. L. Britton, T. J. Stillman, J. Lebbink, W. M. de Vos, F. T. Robb, C. Vetriani, D. Maeder, and D. W. Rice.** 1998. Insights into the molecular basis of thermal stability from the analysis of ion-pair networks in the glutamate dehydrogenase family. Eur J Biochem **255:**336-46.
64. **Vanhove, M., Houba, S., Lamottebrasseur, J., Frere, J.M.** 1995. Probing the determinants of protein stability-Comparison of Class A Beta Lactamases. Biochemical Journal **308:**859-864.
65. **Christendat, D., A. Yee, A. Dharamsi, Y. Kluger, A. Savchenko, J. R. Cort, V. Booth, C. D. Mackereth, V. Saridakis, I. Ekiel, G. Kozlov, K. L. Maxwell, N. Wu, L. P. McIntosh, K. Gehring, M. A. Kennedy, A. R. Davidson, E. F. Pai, M. Gerstein, A. M. Edwards, and C. H. Arrowsmith.** 2000. Structural proteomics of an archaeon. Nat Struct Biol **7:**903-9.
66. **Sadeghi, M., Naderi-Manesh, H., Zarrabi, M., Ranjbar, B.** 2006. Effective factors in thermostability of thermophilic proteins. Biophysical Chemistry **119:**256-270.
67. **Kumar, S., C. J. Tsai, and R. Nussinov.** 2000. Factors enhancing protein thermostability. Protein Eng **13:**179-91.
68. **Karshikoff, A., and R. Ladenstein.** 2001. Ion pairs and the

thermotolerance of proteins from hyperthermophiles: a "traffic rule" for hot roads. Trends Biochem Sci **26:**550-6.

69. **Frankenberg, N., Welker, c., Jaenicke, R.** 1999. Does the elimination of ion pairs affect the thermal stability of cold shock protein from the hyperthermophilic bacterium Thermotoga maritima? FEBS Lett. **454:**299-301.
70. **Grimsley, G. R., Shaw, K.L., Fee, L.R., Alston, R.W., Huyghues-Despointes, B.M., Thurlkill, R.L., Scholtz, J.M., Pace, C.N.** 1999. Increasing protein stability by altering long-range coulombic interactions. Prot. Science **8:**1843-1849.
71. **Szilagyi, A., and P. Zavodszky.** 2000. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. Structure **8:**493-504.
72. **Bosshard, H. R., Marti, D.N., Jelesarov, I.** 2004. Protein stabilization by salt bridges: concepts, experimental approaches and clarification of some misunderstandings J. Mol. Recognit. **17:**1-16.
73. **Honig, B., Hubell, W. L.** 1984. Stability of salt-bridges in membrane proteins Proc. Natl. Acad. Sci. U.S.A. **81:**5412-5416.
74. **Loladze, V. V., Ibarra-Molero, B., Sanchez-Ruiz, J. M., Makhatadze, G. I.** 1999. Engineering a thermostable protein via optimization of charge–charge interactions on the protein surface. Biochemistry **38:**16149-16423.
75. **Wang, X., X. He, S. Yang, X. An, W. Chang, and D. Liang.** 2003. Structural basis for thermostability of beta-glycosidase from the thermophilic eubacterium Thermus nonproteolyticus HG102. J Bacteriol **185:**4248-55.
76. **Makhatadze, G. I., Loladze, V.V., Ermolenko, D.N., Chen, X.F.; Thomas, S.T.** 2003. Contribution of Surface Salt Bridges to Protein Stability: Guidelines for Protein Engineering. J. Mol. Biol. **327:**1135-1148.
77. **Sarakatsannis, J. N., Duan, Y.** 2005. Statistical characterization of salt bridges in proteins. Proteins **60:**732-739.
78. **Marqusee, S., Sauer, R.T.** 1994. Contributions of a hydrogen-bond salt bridge network to the stability of secondary and tertiary structure in lambda-repressor. Prot. Science **3:**2217-2225.
79. **Fukuchi, S., and K. Nishikawa.** 2001. Protein surface amino acid compositions distinctively differ between thermophilic and mesophilic bacteria. J Mol Biol **309:**835-43.

80. **Cambillau, C., and J. M. Claverie.** 2000. Structural and genomic correlates of hyperthermostability. J Biol Chem **275:**32383-6.
81. **Jaenicke, R., and G. Bohm.** 1998. The stability of proteins in extreme environments. Curr Opin Struct Biol **8:**738-48.
82. **Matthews, B. W., H. Nicholson, and W. J. Becktel.** 1987. Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. Proc Natl Acad Sci U S A **84:**6663-7.
83. **Bogin, O., M. Peretz, Y. Hacham, Y. Korkhin, F. Frolow, A. J. Kalb, and Y. Burstein.** 1998. Enhanced thermal stability of Clostridium beijerinckii alcohol dehydrogenase after strategic substitution of amino acid residues with prolines from the homologous thermophilic Thermoanaerobacter brockii alcohol dehydrogenase. Protein Sci **7:**1156-63.
84. **Kannan, N., and S. Vishveshwara.** 2000. Aromatic clusters: a determinant of thermal stability of thermophilic proteins. Protein Eng **13:**753-61.
85. **Zeldovich, K. B., I. N. Berezovsky, and E. I. Shakhnovich.** 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol **3:**e5.
86. **Farias, S. T., and M. C. Bonato.** 2003. Preferred amino acids and thermostability. Genet Mol Res **2:**383-93.
87. **Shimizu, Y., A. Inoue, Y. Tomari, T. Suzuki, T. Yokogawa, K. Nishikawa, and T. Ueda.** 2001. Cell-free translation reconstituted with purified components. Nat Biotechnol **19:**751-5.
88. **Andrade, M. A., S. I. O'Donoghue, and B. Rost.** 1998. Adaptation of protein surfaces to subcellular location. J Mol Biol **276:**517-25.
89. **Pechkova, E., V. Sivozhelezov, and C. Nicolini.** 2007. Protein thermal stability: the role of protein structure and aqueous environment. Arch Biochem Biophys **466:**40-8.
90. **Warren, G. L., and G. A. Petsko.** 1995. Composition analysis of alpha-helices in thermophilic organisms. Protein Eng **8:**905-13.
91. **Higgins, C. L., J. Meyer, and P. Wittung-Stafshede.** 2002. Exceptional stability of a [2Fe-2S] ferredoxin from hyperthermophilic bacterium Aquifex aeolicus. Biochim Biophys Acta **1599:**82-9.
92. **Nakka, M., R. B. Iyer, and L. G. Bachas.** 2006. Intersubunit disulfide interactions play a critical role in maintaining the thermostability of glucose-6-phosphate dehydrogenase from the hyperthermophilic bacterium Aquifex aeolicus. Protein J **25:**17-21.

93. **Volkin, D. B., and A. M. Klibanov.** 1987. Thermal destruction processes in proteins involving cystine residues. J Biol Chem **262:**2945-50.
94. **Tanner, J. J., R. M. Hecht, and K. L. Krause.** 1996. Determinants of enzyme thermostability observed in the molecular structure of Thermus aquaticus D-glyceraldehyde-3-phosphate dehydrogenase at 25 Angstroms Resolution. Biochemistry **35:**2597-609.
95. **Muir, J. M., R. J. Russell, D. W. Hough, and M. J. Danson.** 1995. Citrate synthase from the hyperthermophilic Archaeon, Pyrococcus furiosus. Protein Eng **8:**583-92.
96. **Scandurra, R., V. Consalvi, R. Chiaraluce, L. Politi, and P. C. Engel.** 1998. Protein thermostability in extremophiles. Biochimie **80:**933-41.
97. **Rumfeldt, J. A., C. Galvagnion, K. A. Vassall, and E. M. Meiering.** 2008. Conformational stability and folding mechanisms of dimeric proteins. Prog Biophys Mol Biol **98:**61-84.
98. **Neet, K. E., and D. E. Timm.** 1994. Conformational stability of dimeric proteins: quantitative studies by equilibrium denaturation. Protein Sci **3:**2167-74.
99. **Ponstingl, H., T. Kabir, D. Gorse, and J. M. Thornton.** 2005. Morphological aspects of oligomeric protein structures. Prog Biophys Mol Biol **89:**9-35.
100. **Lukatsky, D. B., B. E. Shakhnovich, J. Mintseris, and E. I. Shakhnovich.** 2007. Structural similarity enhances interaction propensity of proteins. J Mol Biol **365:**1596-606.
101. **Jaenicke, R.** 1991. Protein stability and molecular adaptation to extreme conditions. Eur J Biochem **202:**715-28.
102. **Takai, K., Y. Sako, A. Uchida, and Y. Ishida.** 1997. Extremely thermostable phosphoenolpyruvate carboxylase from an extreme thermophile, Rhodothermus obamensis. J Biochem **122:**32-40.
103. **Lee, D. W., E. A. Choe, S. B. Kim, S. H. Eom, Y. H. Hong, S. J. Lee, H. S. Lee, D. Y. Lee, and Y. R. Pyun.** 2005. Distinct metal dependence for catalytic and structural functions in the L-arabinose isomerases from the mesophilic Bacillus halodurans and the thermophilic Geobacillus stearothermophilus. Arch Biochem Biophys **434:**333-43.
104. **Lee, D. W., Y. H. Hong, E. A. Choe, S. J. Lee, S. B. Kim, H. S. Lee, J. W. Oh, H. H. Shin, and Y. R. Pyun.** 2005. A thermodynamic study of mesophilic, thermophilic, and hyperthermophilic L-arabinose

isomerases: the effects of divalent metal ions on protein stability at elevated temperatures. FEBS Lett **579:**1261-6.

This chapter was partially published in:

Vesna Prosinecki, Patrícia F.N. Faísca and Cláudio M. Gomes, Conformational States and Protein Stability in a Proteomic Perspective, Current Proteomics, Vol. 4 Issue 1, 2007, 44-52

Chapter 2

METHODOLOGIES FOR PROTEOMICS STUDIES ON PROTEIN STABILITY

Methodologies for proteomics studies of protein stability

2.1. Introduction

Results of the work represented in this thesis rely on the use of the various scientific methods whose complementary information provided additional answers or conclusions regarding determinants of proteins stability. The unified approach of several scientific methodologies was chosen in order to address structure, function, folding and stability of proteins, importance in cellular life preserving processes or evolutionary correlations. Structural classification of proteins may provide a most general reference to possible biochemical function and together with bioinformatics approaches insight into evolutionary relationships and consequent stability determinants. In our studies choices of complementing methodologies have been made depending on the feature investigated in order to assess these issues. The objective of the chapter representing methodologies' overview is not a comprehensive survey of protein(s) analysis methods, but to represent state of the art scientific progress in this area that guided scientific judgment, and also to cover the group of methods and approaches that were used in accomplishing this thesis and reasons of using them.

2.2. Identification and quantification

In the study of determinants of protein thermostability, proteomics approach and its methodologies have given us vast range of valuable data. Generally speaking, proteome and subproteome analyses can be broadly categorized into three types of studies: quantitative protein profile comparisons, analysis of protein-protein interactions, and compositional analysis of simple proteomes or sub-proteomes such as organelles or large protein complexes. The experimental global analysis of conformational changes and assessment of protein disorder at a whole proteome scale has become recently facilitated with the development of two novel methodologies that are overviewed in this section.

Profiling hyperstable proteins at a proteomic scale

Proteins with a very high conformational stability can be labeled as hyperstable. Profiling these proteins at a proteomic level is of interest within numerous perspectives. Hyper-stable proteins can be used, for example, in processes relying on biological catalysis (1, 2). Also, the profiling of proteins of a given pathogen according to their stabilities is a potentially valuable approach, as this could allow the identification of potential therapeutic targets. Hyperstable proteins also constitute excellent working models in more fundamental studies aiming at the elucidation of the molecular determinants of the stability of a particular fold (3, 4), or as targets for protein structure determination.

Proteins with Extreme Thermal and Chemical Resistance

This thesis represents a novel approach aimed at profiling a proteome for its most intrinsically stable proteins (5). The methodology was implemented on the proteome of a hyper-thermophile, as the high optimal growth temperatures of these organisms (>80°C) make them valuable sources of proteins with intrinsically high stability. Under such extreme conditions, the stability of the proteome must rely on the combined effect of several factors, in which intrinsic molecular determinants are pivotal. However, the methodology can be applied to any proteome. The outlined procedure consists on thermal and chemical perturbation of the proteome under study as a function of time. The impact of the destabilization during the perturbation time is assessed by periodic sampling, fractionation and electrophoretic analysis. A comparison of the chromatographic and electrophoretic profiles allows the identification of a set of surviving proteins, which are subsequently identified by mass spectrometry methods. Whenever the biological activity of the identified proteins is known, activity assays are performed after the incubation and along the perturbation period to verify if the selected protein retains biological function, a further indication of structural maintenance.

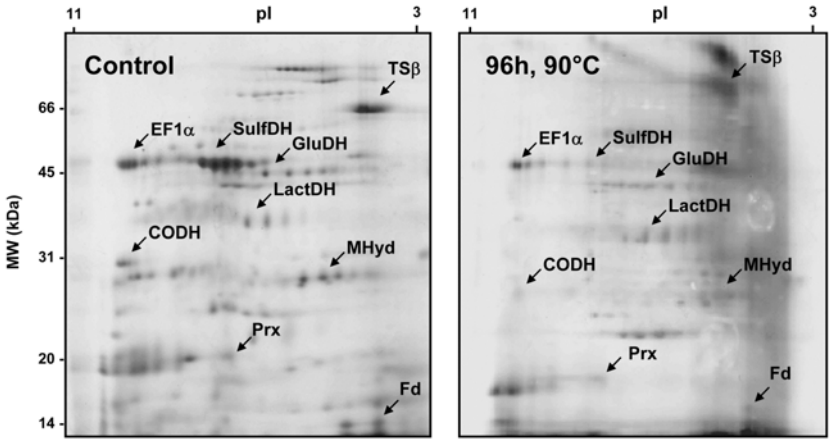


Fig. 2.1. 2DE Silver stained gel (12.5% SDS-PAGE, 500 µg protein), corresponding to the cytosolic proteome of *Sulfurisphaera sp.*, before (A) and after (B) 96 h incubation at 90 °C. Arrows indicate some of the identified proteins: EF1_, elongation factor 1-alpha; SulfDH, sulphide dehydrogenase; TS_, thermosome _-subunit; GluDH, Glutamate dehydrogenase; LactDH, Lactate dehydrogenase; MHyd, metal-dependent hydrolase; Prx, peroxiredoxin; Fd, Ferredoxin [3Fe4S][4Fe4S]. (5) See Chapter Three for further details.

This procedure was employed to profile the cytosolic proteome of the thermoacidophile *Sulfurisphaera* (OGT 85°C). The identification of hyperstable proteins in such a drastic thermophilic background required extreme perturbation protocols relying in extensive thermal perturbation (up to 96h at 90°C) or combined thermal and chemical perturbation (up to 48h at 90°C and 1M GuHCl). The cytosolic fractions perturbed during 96h at 90°C were resolved by two-dimensional electrophoresis (Fig. 2.1): it became clear that the number of proteins decreases substantially and that even after such extensive incubation; there are still a large number of proteins which can be detected. MALDI peptide mass fingerprinting analysis led to the identification of proteins comprised in three distinct functional categories: (a) cellular processes and detoxification, (b) DNA binding, translation and protein modification, and (c) energy metabolism. The proteome perturbation method was validated by verifying the extent to which the subset of proteins obtained after the perturbation

corresponds to biologically active molecules. For this purpose, two of the identified proteins were selected from the perturbed subset for activity assays, namely peroxiredoxin (Prx) and superoxide dismutase (SOD). The activity of these enzymes was measured in the native and perturbed proteome. Both activities were not only present but a specific activity of the enzymes was also observed (Fig. 2.6). This observation is compatible with a significant enrichment of these two enzymes in the thermally treated extract, as a result of their enhanced thermostability.

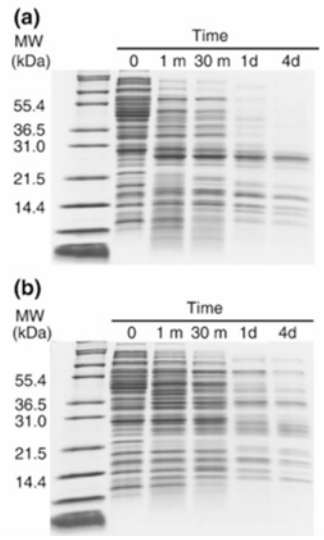


Fig. 2.2. Profiling proteins according to proteolytic resistance. A cell lysate from *E. coli* was digested with trypsin (a) and thermolysin (b) up to four days, and periodically sampled for SDS-PAGE gel analysis. A significant number of proteins were digested within the first 30 min; some proteins, however, the so-called survivors, persisted up to 4 days. (6)

Proteins Resistant to Proteolysis

Recently, Marqusee and co-workers described an investigation on proteolytic susceptibility/resistance of proteins from a proteomics point of view using the *E. coli* proteome as model system (6). The rationale of the approach was that proteolytic susceptibility reflects the accessibility of cleavable states determined by the protein conformation, and not by the overall protein stability. However, it should be noted that *in vivo*, or

in the context of a complex mixture of proteins, proteolysis may be controlled by formation of complexes. In order to profile proteins according to their proteolytic resistance, an assay was developed where proteins originating from the cell lysate were submitted to an extensive proteolytic treatment with trypsin or thermolysin during long period of time. It is clear that digestions made up to 4 days lead to a successive enrichment on the so called survivors (Fig. 2.2.), which were further identified via the mass spectrometry of spots excised from 2DE gels (6). The profiled proteins were then compared according to several criteria among which are amino acid composition, protein folds, and biological function. An important control of this approach arose from the analysis of the amino acid composition of the identified survivors, which demonstrated that proteolytic resistance is not arising from the absence of potential cleavage site residues. Further, this analysis showed that the identified proteins do not share any common structural features that could account for their proteolytic resistance.

In-Gel Detection of Protein Surface Hydrophobicity Changes

Modifications on the conformation of a protein frequently involve a variation on the surface hydrophobicity as a result of the exposure of otherwise solvent-shielded apolar groups. Chaudhuri and co-workers have established a procedure that allows monitoring changes in the exposure of surface hydrophobic domains in proteins as a result of a conformational change (7). In their study they took the advantage of the ability of Bis-ANS (4,4'-dianilino1,1'-binaphthyl-5,5'-disulfonic acid) to interact with the protein and covalently bind to the site of interaction after being exposed to ultraviolet (UV) irradiation. This apolar fluorescent probe binds preferentially to hydrophobic patches, which become solvent accessible as a result of a moderate conformational change. The rationale behind this approach is that the intensity of Bis-ANS fluorescence will depend on the variation of the protein surface hydrophobicity, which can be correlated to a change in the protein conformation. The fact that this molecule can also be photolabeled to targets makes it a useful conformational probe. This approach has the plus of being able to discriminate partially altered conformational states, as shown in validation studies performed on rhodanase (7) (Fig. 2.3);

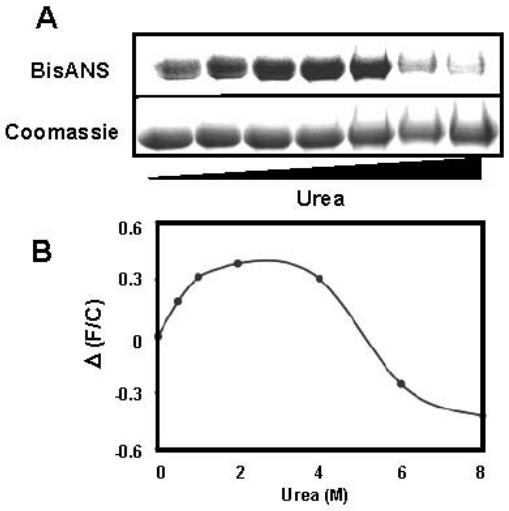


Figure 2.3. In gel detection of changes in protein conformation**.** The known effect of urea on rhodanase surface hydrophobicity was used to validate the Bis-ANS labelling and detection procedure. Rhodanase (1 mg/ml) was incubated with different urea concentration and labelled with 0.1 nM Bis-ANS, and analysed by SDS-PAGE. Densitometry quantitated the Bis-ANS fluorescence (F), which was normalized for the Coomassie (C) staining (A). The dependence of fluorescence intensity _F/C was plotted as a function of urea concentration (B): the maximum at 2 M indicates the formation of stable intermediate complexes with increased surface hydrophobicity. For concentrations of urea above 4 M the protein is likely to be unfolded, as suggested by the dramatic decrease in Bis-ANS incorporation. (7)

However, it does not identify unfolded states and aggregates, as in these cases the hydrophobic regions necessary for the interactions are either absent or unavailable. The ultimate goal of this methodology is its application at a proteomic scale to screen for conformational changes occurring *in vivo* as result of a pathological process or a stress factor. This assay was used to analyze the effect of oxidative stress on a complex mixture of cytosolic proteins originating from skeletal muscle. The proteome was exposed to increasing levels of oxidative stress using either *in vitro* metal-catalysed oxidation, or *in vivo* denervation, a process that induces oxidative stress. Subsequent photoincorporation of Bis-ANS

and resolution of the proteome by 2DE led to the observation that, in both circumstances, a significant variation of the surface hydrophobicity was noted in two major proteins from skeletal muscle proteins, creatine kinase and glyceralaldehyde-3-phosphate dehydrogenase. These two proteins are among those that undergo a conformational change upon oxidative stress, serving therefore to illustrate the validity of the approach (7). While fluorophore labeling approaches have the merit of allowing an evaluation of the conformational change of a single protein within a complex proteome, some aspects should be noted. For example, this approach fails to detect conformational states lacking a substantial amount of the necessary superficial hydrophobic reporter regions, such as aggregates or denatured states. Additionally, the fact that the extent of labeling differs among proteins reflects different propensities towards the probe, which may result in misleading results. Considering these aspects, some of which are intrinsic to the nature of the fluorophore-protein interaction, this methodology elegantly allows a direct analysis of a conformational change in a particular set of proteins, in the background of a complex proteome.

Electrophoretic Detection of Intrinsically Disordered Proteins

The initial discovery of structurally disordered proteins resulted from studies on isolated expressed proteins, whose properties resembled those of the denatured states of globular proteins with respect to structural content, conformational flexibility, and hydrodynamic radius (8). A series of subsequent proteomic approaches based on the use of acid, organic solvents and heat treatment to selectively enrich extracts for disordered proteins have been established. These strategies and respective results, which have been addressed in a recent review (9), rely on the principle that under these harsh conditions globular proteins will precipitate whereas intrinsically disordered protein will remain in solution. The combined use of 2DE and mass spectrometry (MS) has allowed the identification of a series of putative disordered proteins. These approaches constitute valid starting points, but the fact that some apparent positives may comprise proteins with an intrinsic high stability or proteins that do not necessarily precipitate upon denaturation and

have the ability to refold upon cooling, can not be fully discarded.

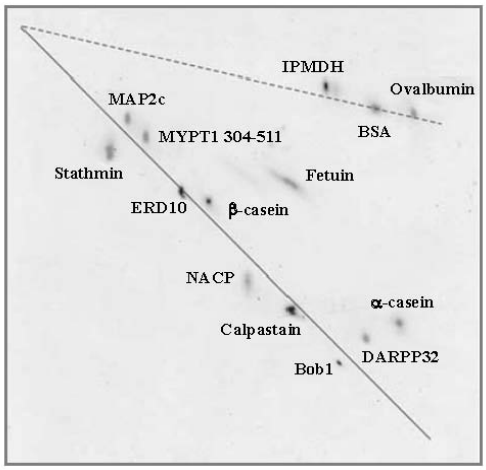


Fig. 2.4. Electrophoretic detection of intrinsically disordered proteins. A mixture of unstructured and globular proteins (1 µg each) was resolved on a native 7.5% gel in the first dimension and on a 7.5% denaturing 8 M urea gel in the second dimension. For experimental details see (Csizmok et al., 2007; Csizmok et al., 2006). Proteins migrating along the solid diagonal line those with disordered structure. Globular proteins: fetuin, IPMDH, BSA, ovalbumin; Disordered proteins: stathmin, MAP2c, MYPT1 304-511, ERD10, _-casein, NACP, CSD1, Bob1, DARPP32, _-casein. (9)

More recently, a novel technique was established by Tompa and colleagues that enables detection of proteins with intrinsic disorder (10). This methodology provides direct evidence for protein structural disorder, thus circumventing some of the above-mentioned limitations of the previous strategies. The outlined methodology results from the combination of native and denaturing electrophoresis of heat-treated samples. The rationale of the approach is that the heat treatment step will simplify the initial protein mixture by removing globular proteins that precipitate under these conditions. The subsequent first dimension native electrophoresis will resolve proteins according to their charge/mass ratios. In the second dimension in 8 M urea, disordered proteins that are unaffected by the denaturant will migrate the same as

in the first dimension, and end along the diagonal of the gel; as a result of an increased friction coefficient heat resistant proteins will migrate less and stay above the diagonal. This methodology was tested and validated using known disordered proteins, and control globular proteins (Fig. 2.4.), and later expanded to the analysis of cellular extracts of *Saccharomyces cervisiae* and *Escherichia coli*. Again, some limitations must be taken into consideration, namely the fact that the native first dimension has a poor resolving power and that basic proteins are lost under standard conditions (Csizmok et al., 2006). Overall, this methodology is an important step forward towards the identification of the "disorderome" (9, 10).

Isobaric tags for relative and absolute quantitation - iTRAQ

A new class of isobaric reagents, developed by Applied Biosystems, Isobaric tags for relative and absolute quantitation iTRAQ™ (11), is used for characterization of protein mixtures, in order to understand complex biological systems. Aim of such research is to monitor changes of proteins in perturbed systems. The iTRAQ™ reagents are a set of isobaric reagents which are specific for amino group and allow identification and quantification of up to four different samples simultaneously. The amine specificity of these reagents, including the N-terminus and the ε-amino group of the lysine side-chain, makes most peptides in samples amenable to this labeling strategy with no loss of information from samples that were subjected to post-translational modifications (e.g. phosphorylation) (12). The multiplexing capacity of iTRAQ reagents allows replication of the information within LC-MS/MS experimental regimes, which provides additional statistical validation within experiment. The enhancement of individual proteins´ signal, which may be in low abundance in any sample, is provided by a set of single unresolved additive precursor ions in MS within the resultant mixture. The four reporter group ions appear as distinct masses between *m/z* 114–117 after collision-induced dissociation (CID) of the parent fragment produced in MS (e.g. MS/MS), while the sequence informative *y*- and *b*- ions remain as additive isobaric signals. Performing studies supported by iTRAQ technology requires the ability to execute certain differential comparison of a given biological mixture state in reference to a control (Fig.2.5).

The results presented in this thesis demonstrate an example of the wide scientific approach based on one hand with the information that were achieved by complementing iTRAQ identification and quantification results, and on the other hand with other complementary experimental, instrumental and calculation methods. iTRAQ was an excellent method of choice as we were following highly stable proteome subsets isolated along the time by extensive thermal treatment and comparing those with the initial untreated proteome sample.

2.3. Bioinformatics

Bioinformatics combines the tools and techniques of mathematics, computer science and biology in order to understand the biological significance of a variety of data. The ultimate goal of the field is to enable the discovery of new biological insights as well as to create a global perspective from which unifying principles in biology can be discerned. The automated classification of proteins into categories according to their sequence, structure, function, evolution, etc. uses various criteria in attempt to systematize available data and give deeper insight.

All the information or data from experimental research are stored in publicly internet available databases, accompanied by softwares for further data analysis and links to similar databases. Every protein database contains defined types of information regarding either sequence or structural data or combination of both.

Swiss-Prot database (http://www.expasy.org/sprot/) is maintained by the Swiss Institute of Bioinformatics in collaboration with European Bioinformatics Institute. It is a sequence database with large number of protein records with bibliographical references including comments on biological function, secondary structure and links to other databases relevant for the particular sequence, including some additional comments.

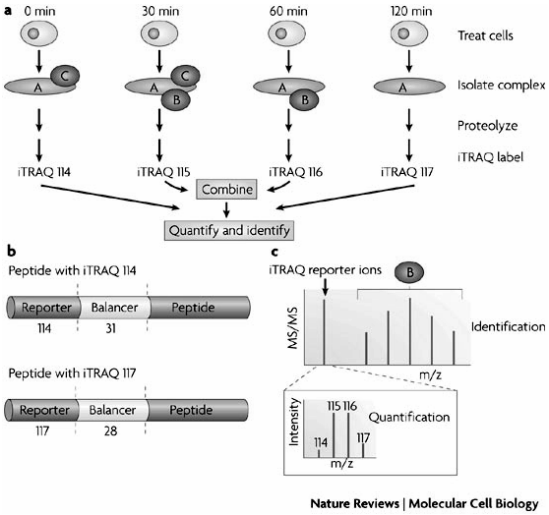


Fig. 2.5 a) Desired treatment of cells is followed by isolation of protein complexes and proteolysis. Isobaric tags (iTRAQ) are chemically added to the N terminus of every peptide (as well as to lysine ε-amine groups). Samples from multiple treatment time points are combined and subjected to analysis. b) A peptide labeled with the iTRAQ 114 and iTRAQ 117 reagents. iTRAQ is isobaric, such that addition of the 114Da or 117Da mass tags alter the mass of a given peptide by the same amount. To maintain a constant mass, the reporter moiety (for example, of mass 114) is separated from the peptide by a balancer group. The reporter and balancer groups fragment in the collision cell of the mass spectrometer during the tandem mass spectrometry (MS/MS) event, and the intensity of the reporter ions is monitored. c) Analysis of an iTRAQ experiment. MS/MS analysis of a labeled peptide generates a fragmentation spectrum that yields the sequence of the peptide. The iTRAQ reagent is fragmented in the same step and reporter ions are quantified by magnifying the low mass range (114–117) area. In the example shown, protein B associates with protein A (the bait) after 30 and 60 minutes of stimulation, but not after 120 minutes of treatment. m/z, mass/charge ratio. (13)

PROSITE database (http://www.expasy.org/prosite/) is in close relationship with Swiss-Prot protein sequence data bank. It contains information regarding biologically significant primary structure patterns that characterize protein families with same functions. It consists of

patterns and profiles in order to classify protein sequences or its domains to a known protein family. For each family, details are available regarding the structure and function of corresponding proteins, giving the information concerning protein homology. Each entry is fully documented including description of the family and the reasons leading to a development of the profile or pattern. Profiles or patterns that are chosen are specific enough to avoid detecting too many unrelated sequences, but still to detect the most or all of the relevant members of the family. Each entry is periodically reviewed to check its validity, updates are regular, and software tools that are available on the PROSITE are used to automatically update relevant Swiss-Prot entries.

Protein Data Bank PDB (http://www.rcsb.org/pdb/) is the major structure database consisting of records of experimentally determined 3D structures, including references. All of the records contain detailed structural data from X-ray crystallography or NMR spectroscopy: atomic coordinates, primary structure residues, and secondary structure residues, all of them being evaluated for its accuracy by specialized software before releasing information for public use. Each record has unique identifier.

Class Architecture Topology Homology database CATH (http://www.biochem.ucl.ac.uk/bsm/cath_new/index.htlm) consists of hierarchical classification of independent structural domains of PDB deposited sequences that have resolution equal or less than 3 A. Proteins consisting of more than one domain are analyzed by automated algorithms for domain recognition or by arguments in the literature. Levels in CATH classification are class, architecture, topology – fold family and homologous superfamily, based on the elements of the secondary structure. Automated algorithms for domains recognition are used for analysis of proteins with more than one domain. The classification is based on the secondary structure and grouped in to four categories: mainly alpha – majority of proteins secondary structure are α helices; mainly beta – majority of secondary structure are extended β strands; alpha-beta – with α/β and $\alpha+\beta$ structures and various formations of secondary structure as the fourth group. Interconnections (like e.g.

barrels) are not included. Both orientation and interconnection are used for the topology classification. Sequence similarity is the criteria for organizing secondary structure into the homologous superfamilies level.

Structural Classification of Proteins SCOP (http://scop.mrc-lmb.cam.ac.uk/scop/) database provides structural description with evolutionary relationships for the proteins with known structure stored in PDB (14). SCOP provides information regarding atomic coordinates, images of the structure, sequence, conformational changes related to the function and literature references for all the known structures. The classification levels are the family, superfamily, fold and the class. At the family level the sequence similarity is 30%, with some exceptions. At superfamily level proteins with low sequence similarity are classified together but their structure and function indicate a probable common ancestor. At the fold level proteins are grouped according to the same way in terms of orientation and topological connections. At the class level, four groups are formed according to the secondary structure: all α, all β, α/β – α helices and β strands alternate in the structure, and $\alpha+\beta$ – α helices and β strands are found in the distinct structural regions.

All these databases facilitate structural comparison and provide better understanding of structure and function. In spite of the complex structure and random length, proteins with similar structure/function often share a common ancestor and similar amino acid sequence. Evolution has gradually influenced the sequence change by substitution, insertion or deletion, so grouping into families of common function may be accomplished by comparing and aligning sequences. Besides from that, the proteins sharing similar function should share similar 3D structure. These homolog proteins have evolutionary relationship.

2.4. Preservation of protein structure in solution

Within a cell, conformational changes in a protein may result from a functional modification, for example, as a result of an interaction with an activity modulator (such as another protein, a nucleic acid or a metal ion) or they may reflect the onset of more drastic alterations, such as misfolding, degradation or aggregation. Unlike analysis of complex mixtures of proteins, in pure protein solutions, these alterations can be monitored

using different biophysical methods such as circular dichroism (CD), fluorescence, nuclear magnetic resonance (NMR), infrared spectroscopies (IR), dynamic light scattering (DLS) and many other approaches (15, 16). Also, alterations in protein structure may influence its biological activity; therefore biochemical methods including activity assays are able to detect presence of structural modification.

In this thesis, proteome thermal perturbation method (presented in Chapter Three) was validated by verifying the extent to which the subset of proteins obtained after the perturbation corresponds to biologically active molecules. For this purpose, some of the identified proteins were selected from the thermally perturbed subset. Ferredoxins are widespread in the three domains of life which makes them exelent investigation models. They are small, monomeric peptides containing iron–sulfur centers whose integrity can be easily followed by several spectroscopic techniques. Folded structure of ferredoxin purified from the thermally treated proteome was tested with fluorescence spectroscopy measurements, monitoring tryptophan fluorescence emission between 300-450 nm, with excitation at 280 nm. A preliminary biophysical characterization was carried out on this purified band, and its assignment as a Fd was corroborated by obtaining the typical UV-visible spectrum, with the characteristic band at 410 nm corresponding to native Fd with intact Fe-S clusters (Fig. 2.7). These proteins are characterized by a very high thermal stability (17) with very high midpoint melting transitions (Tm 110°C at pH 7 for the di-clusters, seven-iron-containing ferredoxins from the archaeal organism *Acidianus ambivalens*) as well as a very high kinetic stability (18) The protein core harboring the two iron-sulfur centers is likely to play an important role in protein thermostability, as well as a Zn-containing N-terminal extension (18, 19).

Thermophilic proteins are intrinsically stable and exhibit a high kinetic stability towards thermal degradation. We have validated our proteome perturbation method by verifying the extent to which the subset of proteins obtained after the perturbation corresponds to biologically active molecules. For activity assays we have chosen peroxiredoxin Bacterioferritin co-migratory protein (Prx-BCP) and superoxide dismutase

(SOD), enzymes that are respectively involved in the detoxification of peroxides and superoxide.

The activity of these enzymes was measured in the native and perturbed proteome. Both activities were not only present but a specific activity of the enzymes was also observed (Fig. 2.6). This observation is compatible with a significant enrichment of these two enzymes in the thermally treated extract, as a result of their enhanced thermostability.

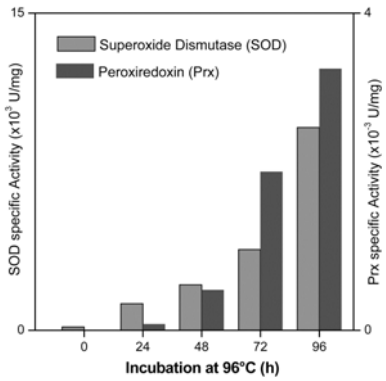


Fig. 2.6. Activity profiles of oxidative stress proteins during thermal perturbation. The specific activity of superoxide dismutase (SOD) and peroxiredoxin (Prx) were measured as function of incubation at 90 °C. The increase in the specific activity shows that upon 96 h perturbation the proteins retain their biological function and presumably their native fold (5). See Chapter Three for further details.

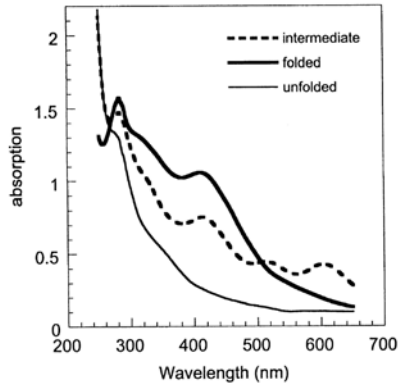


Fig. 2.7. *Uv*-Visible absorption of folded *(*solid thick line*)*, intermediate *(*dashed thick line*)* and unfolded *(*thin solid line*)* ferredoxin at pH 10. Unfolding was promoted by addition of 7 M GuHCl *(*pH 10*)* to the protein (17).

2.5. Conclusion

The methodologies discussed here illustrate efforts to snapshot the conformational state and structure of individual proteins within a complex proteome and also relates to the recent attempts to profile proteins at a proteomic scale according to their enhanced conformational stability, which suggests that some metabolic and cellular processes may require particularly stable proteins. This hypothesis has been originally put forward following the profiling of our chosen archaeal model organism, *Sulfurisphaera*, and its 'stabilome': the identified proteins participate in cellular processes (e.g.. defense against reactive oxygen species, nucleic acid protection and energy production) in which some key proteins have enhanced thermal stabilities, which may relate to their importance on the metabolism of the organism, as it will be further discussed. Overall, the reviewed approaches pave way to further proteome-level studies aiming at characterizing the structural and conformational properties of proteins in a proteomic context. Future challenges encompass framing this knowledge in the context of the cell physiology, regulation of protein function and metabolic networks.

2.6. References

1. **Cowan, D. A.** 1992. Biotechnology of the Archaea. Trends Biotechnol **10:**315-23.
2. **Li, W. F., X. X. Zhou, and P. Lu.** 2005. Structural features of thermozymes. Biotechnol Adv **23:**271-81.
3. **Fitter, J.** 2005. Structural and dynamical features contributing to thermostability in alpha-amylases. Cell Mol Life Sci **62:**1925-37.
4. **Perl, D., and F. X. Schmid.** 2001. Electrostatic stabilization of a thermophilic cold shock protein. J Mol Biol **313:**343-57.
5. **Prosinecki, V., H. M. Botelho, S. Francese, G. Mastrobuoni, G. Moneti, T. Urich, A. Kletzin, and C. M. Gomes.** 2006. A proteomic approach toward the selection of proteins with enhanced intrinsic conformational stability. J Proteome Res **5:**2720-6.
6. **Park, C., S. Zhou, J. Gilmore, and S. Marqusee.** 2007. Energetics-based protein profiling on a proteomic scale: identification of proteins resistant to proteolysis. J Mol Biol **368:**1426-37.

7. **Pierce, A., E. deWaal, H. Van Remmen, A. Richardson, and A. Chaudhuri.** 2006. A novel approach for screening the proteome for changes in protein conformation. Biochemistry **45:**3077-85.
8. **Tompa, P.** 2002. Intrinsically unstructured proteins. Trends Biochem Sci **27:**527-33.
9. **Csizmok, V., Z. Dosztanyi, I. Simon, and P. Tompa.** 2007. Towards proteomic approaches for the identification of structural disorder. Curr Protein Pept Sci **8:**173-9.
10. **Csizmok, V., E. Szollosi, P. Friedrich, and P. Tompa.** 2006. A novel two-dimensional electrophoresis technique for the identification of intrinsically unstructured proteins. Mol Cell Proteomics **5:**265-73.
11. **Ross, P. L., Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin.** 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics **3:**1154-69.
12. **Zieske, L. R.** 2006. A perspective on the use of iTRAQ reagent technology for protein complex and profiling studies. J Exp Bot **57:**1501-8.
13. **Gingras, A. C., M. Gstaiger, B. Raught, and R. Aebersold.** 2007. Analysis of protein complexes using mass spectrometry. Nat Rev Mol Cell Biol **8:**645-54.
14. **Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia.** 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol **247:**536-40.
15. **Plaxco, K. W., and C. M. Dobson.** 1996. Time-resolved biophysical methods in the study of protein folding. Curr Opin Struct Biol **6:**630-6.
16. **Schmid, F. X.** 2005. Spectroscopic techniques to study protein folding and stability. In protein folding handbook p. 22-43. *In* a. T. K. J. Buchner (ed.), vol. 1. Wiley-VCH, Darmstadt.
17. **Wittung-Stafshede, P., C. M. Gomes, and M. Teixeira.** 2000. Stability and folding of the ferredoxin from the hyperthermophilic archaeon Acidianus ambivalens. J Inorg Biochem **78:**35-41.
18. **Gomes, C. F., A.; Carita, J.; Mendes, J.; Regalla, M.; Chicau, P.; Huber, H.; Stetter, K.O.; Teixeira, M.** 1998. Di-cluster, seven-iron ferredoxins from hyperthermophilic Sulfolobales. JBIC **3 (1):**499-507.

19. **Leal, S. S., and C. M. Gomes.** 2005. Linear three-iron centres are unlikely cluster degradation intermediates during unfolding of iron-sulfur proteins. Biol Chem **386:**1295-300.
20. **Niwa, T., B. W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, and H. Taguchi.** 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. Proc Natl Acad Sci U S A **106:**4201-6.

This chapter was published in:

Vesna Prosinecki, Hugo M. Botelho, Simona Francese, Guido Mastrobuoni, Gloriano Moneti, Tim Urich, Arnulf Kletzin and Cláudio M. Gomes, A Proteomic Approach toward the Selection of Proteins with Enhanced Intrinsic Conformational Stability, Journal of Proteome Research**,** Vol. 5, No. 10, 2006, 2720-2726

Note:

This work is done in collaboration with Mass Spectrometry Centre, University of Florence, Florence, Italy, and Institute of Microbiology and Genetics, Darmstadt, Germany. SA, GM and GM performed Mass Spectrometry experiments; TU and AK provided cell extracts of *Sulfurisphaera sp.*

M. Regalla (Amino Acid Sequencing Service, ITQB) has performed N-terminal analysis.

Chapter 3

A PROTEOMICS APPROACH TOWARDS THE SELECTION OF PROTEINS WITH ENHANCED INTRINSIC CONFORMATIONAL STABILITY

A proteomic approach towards the selection of proteins with enhanced intrinsic conformational stability

3.1. Summary

A detailed understanding of the molecular basis of protein folding and stability determinants partly relies on the study of proteins with enhanced conformational stability properties, such as those from thermophilic organisms. In this study we set up a methodology aiming at identifying the subset of cytosolic hyperstable proteins using *Sulfurisphaera sp.*, a hyperthermophilic archaeon, able to grow between 70-97°C, as a model organism. We have thermally and chemically perturbed the cytosolic proteome as a function of time (up to 96h incubation at 90°C), and proceeded with analysis of the remaining proteins by combining one and two dimensional gel electrophoresis, liquid chromatography fractionation, and protein identification by N-terminal sequencing and mass spectrometry methods. A total of 14 proteins with enhanced stabilities which are involved in key cellular processes such as detoxification, nucleic acid processing and energy metabolism were identified including a superoxide dismutase, a peroxiredoxin and a ferredoxin. We demonstrate that these proteins are biologically active after extensive thermal treatment of the proteome. The relevance of these and other targets is discussed in terms of the organism's ecology. This work thus illustrates an experimental approach aimed at mining a proteome for hyperstable proteins, a valuable tool for target selection in protein stability and structural studies.

3.2. Introduction

Thermophilic organisms are a valuable source of proteins with very high stability, the so-called hyperstable proteins. At working temperatures above 80°C the stability of the proteome is kept due to the combined effect of several factors (1, 2). Among these, extrinsic factors like the compatible solutes, acting as chemical stabilisers such as trehalose and β–mannosylglycerate are particularly relevant (3). The concentration of these molecules in the cell can be very high thus resulting in a stabilizing

effect which adds to intracellular protein crowding effects. However, intrinsic factors play a key role in accounting for the high thermal stability of thermophilic proteins. Protein stability relies essentially on the additive effect of non covalent bonds of weak magnitude (4), which result in an overall stabilization of the folded native conformation. Although there is not a general rule that accounts for the molecular determinants, a series of particularities have been outlined in proteins from thermophiles which could explain their increased stability (1, 2). Surface ion pairs, a solvent exposed hydrophobic surface and efficient anchoring of terminal extensions and loop regions seem to be particularly relevant in enhancing protein stability. Thermophilic proteins have an emerging interest both from a fundamental and an applicative point of view. In particular thermophilic enzymes are of interest due to their potential application in biotechnological or industrial processes where a biological catalyst is used (5, 6). Moreover thermophilic proteins constitute also excellent working models in more fundamental studies aiming at the elucidation of the molecular determinants of a particular fold (7-9) or protein structure e.g. (6, 10). This knowledge is essential for the understanding of folding mechanisms and contributes to the design of better structure prediction algorithms.

Following our long standing interest in the study of the stability properties of thermophilic proteins e.g. (8, 11-18), we set up a methodology which aims at identifying a subset of the proteome of thermophilic organisms containing the most intrinsically stable proteins. The rational for our approach is based on previous reports of thermophilic proteins which at physiological pH conditions (pH 6-7) remain folded and functional even upon extensive incubation periods at high temperatures e.g.(11). The thermoacidophilic archaea *Sulfurisphaera* (19) was selected as a model: it is a facultative anaerobe, living in acidic, hot, solfataric springs at high temperatures with an optimal growth temperature of 85^{o} C, tolerating a broad pH range from 1 to 5.5, with an optimal pH at 2. Its genome, which is currently being sequenced, is more than 90% identical to those of the closely related organism *Sulfolobus tokodaii* (20), for which there is inclusively proteomic data available (21). Its broad growth temperature range (70-

97^{o}C) allowed the study and analysis of the proteome obtained from cells grown close to physiological growth temperature extremes. We have outlined a protocol in which the cytosolic proteome of *Sulfurisphaera* was chemically and thermally perturbed as a function of time. The impact of the destabilisation was assessed by combining several approaches including liquid chromatography, SDS-PAGE, two dimensional gel electrophoresis (2DE), N-terminal sequencing and mass spectrometry methods for the identification of purified proteins and spots of interest. A set of proteins with enhanced stabilities was identified and their relevance on a thermophilic background is discussed.

3.3. Experimental

Cell mass and preparation of the cytosolic extract

Sulfurisphaera sp. cells were grown at 72°C (SulfCP72) and 92°C (SulfCP92) according to published procedures(22). Preparation of the cellular fractions was essentially performed as described (23) using 40 mM phosphate buffer, pH 6.5 throughout all stages. All steps were carried out at 4°C in the presence of the protease inhibitor PMSF (0.5 mM). The soluble extracts were divided in small portions and stored at -20°C prior to use.

Thermal and chemical perturbation protocols

The soluble extracts from SulfCP72and SulfCP92 were perturbed by two different protocols in 40 mM phosphate buffer, pH 6.5. The thermal perturbation included incubation at 90°C for up to 96 h and sampling every 24 h. In the second perturbation protocol 1 M of the chemical denaturant guanidinium hydrochloride (GuHCl) was added to the samples, with subsequent incubation periods at 90°C. In both cases, after sampling, less stable precipitated proteins were removed by 10 min centrifugation at 14000 rpm on a bench centrifuge at 4°C.

Liquid chromatography analysis of perturbed proteomes

Soluble proteomes obtained by both perturbation protocols were chromatographically resolved using anionic exchange. A HiTrap Q-sepharose FF (5ml) column equilibrated with 40mM phosphate buffer pH 6.5 was used, and fractions eluted with a linear gradient up to 1M NaCl.

Separated peaks were concentrated (Amicon, 5kDa) and loaded into a gel filtration column (Superdex-75). Protein profiles were obtained monitoring absorption at 280nm. Eluted fractions were checked by 12.5% SDS-PAGE and UV-visible spectrum.

2D electrophoresis

For 2D-PAGE samples were separated in the first dimension by isoelectrofocusing (IEF). Samples of thermally treated as well as not treated soluble extract were precipitated over night with 10% tricloroacetic acid (TCA) in acetone, pellet was additionally washed with acetone and left to air-dry before solubilizing in immobilized pH gradient IPG rehydration buffer composed of 8 M Urea, 2%(w/v) CHAPS, 60 mM DTT and 0.5% IPG buffer (pH 3-11, Amersham Biosciences). Samples were rehydrated using 13 cm long pH 3-11 NL IPG stripes (Amersham Biosciences). Isoelectric focusing was carried out using the following scheme: 30 V for 12 h, 250 V for 250 Vhr, 500 V for 750 Vhr, 1000 V for 1500 Vhr, 2500 V for 2500 Vhr and 8000 V for 3237 Vhr, and finally 8000 V for 24000 Vhr. The stripes were equilibrated for 15 min in 5 ml solution per stripe containing 50 mM Tris-HCl pH 8.8, 6 M Urea, 30% (v/v) glycerol, 2% (w/v) SDS, with the addition of 0.1% (w/v) DTT and 0.25% (w/v) iodoacetamide respectively in the same solution composition. The equilibrated IPG stripes were placed on the top of the 12.5% SDS gel, sealed with 0.5% (w/v) agarose and run for 15 min on 30 mA, followed by 30 mA per gel until the end of the run. Silver staining was performed with “Silver staining kit – protein” (Amersham Biosciences), and gel images were analyzed using ImageMaster 2D Platinum (Amersham Biosciences).

In situ gel digestion

Complete Coomassie de-staining of excised gel bands/spots was performed by shrinking and rehydratating the gel pieces with acetonitrile and 50 mM ammonium bicarbonate respectively. After vacuum drying in Speedvac (Thermo, San José, CA), the excised bands were rehydratated and submitted firstly to chemical reduction in a solution of DTT 10 mM in 50 mM ammonium bicarbonate left for 30 min at 56°C, and then to alkylation with a iodoacetamide solution in 50 mM ammonium

bicarbonate in the dark, for 30 min, at room temperature. *In situ* digestion was performed on both bands and spots by rehydrating the gel plugs with a solution of 12 ng/µl and 6 ng/µl respectively, of modified trypsin (Promega, Madison, WI) in 10 mM ammonium bicarbonate solution and left to incubate for 90 min at 4°C. Supernatants were recovered, blocked by adding TFA 10% and spotted on an AnchorChip™ (Bruker Daltonics, Bremen, Germany) target plate. 1 µl of sample was deposited on the target plate and allowed to dry; 0.35 µl of matrix (5 g/l alpha-cyano-4-hydroxycinnamic acid in 50/50 acetonitrile/0.1% TFA) were then added and, again, allowed to dry. To the remaining gel plugs, 30 µl of 10 mM ammonium bicarbonate were added and the reaction proceeded overnight at 37°C. Supernatants were eventually withdrawn and analysed by MALDI MS and MS/MS.

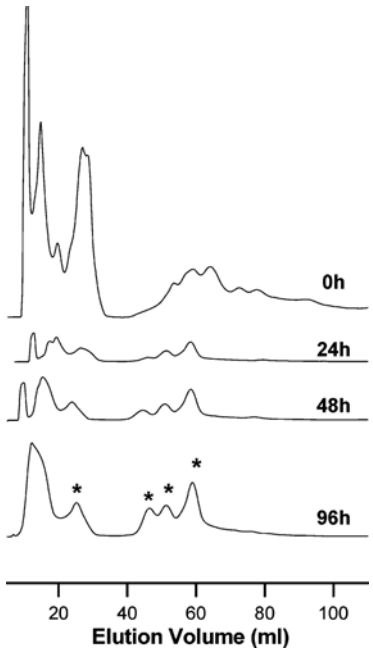


Fig.3.1 Protein chromatograms of purification of SulfCP72 thermally treated proteome. The extract is being selectively enriched in thermostable proteins (marked with an asterisk), which elute at identical ionic strengths despite the longer incubation at 90 °C. In agreement, a decreasing number of bands were observed in denaturing electrophoresis (see Fig. 3.2)

MALDI Peptide mass fingerprinting

Mass spectrometry analysis was performed on a matrix assisted laser desorption ionization tandem mass spectrometer having time-of-flight/time-of-flight optics (Ultraflex MALDI-TOF/TOF, Bruker Daltonics) by using the data acquisition software FlexControl™ 2.4. Mass spectra were acquired in reflectron mode over the m/z range 800-3500. The instrumental parameters were chosen by setting the Ion source 1 at 25 kV, the reflector at 26.30 kV and the delay time at 20 ns. The instrument was externally calibrated prior to analysis by using the Bruker peptide calibrant kit (1000-3000 Da) and the sample spectra internally recalibrated with trypsin autolysis signals. Mass spectra were elaborated by using FlexAnalysis™ 2.4. The peptide masses present in each mass spectrum, through the integrated software Biotools™ 2.2., are used to search the NCBI non-redundant database by using MASCOT software available on line (http://www.matrixscience.com/cgi/nph-mascot.exe?1) which compares the experimentally determined tryptic peptide masses with theoretical ones calculated for proteins contained in the protein database. By using NCBInr as protein databank to search against, Mascot significance threshold is 64. The taxonomy was set on Archea; carbamidomethylation and methionine oxidation were selected as complete and partial modifications respectively; two missed cleavages were allowed for trypsin chosen as cleaving agent. Searches were performed setting a mass tolerance ranging from 20 to 40 ppm. Quick confirmation of our results was made by searching also against SWISS-PROT http://expasy.org/sprot database by using the softwares Profound http://65.219.84.5/service/prowl/profound/profound_E_adv.html and MS-FIT http://prospector.ucsf.edu/ucsfhtml4.0/msfit.htm which use different searching algorithms to identify proteins. The accession numbers were retrieved from the NCBI http://www.ncbi.nlm.nih.gov.

MALDI MS/MS peptide sequencing

The best s/n peptides signals present in each MS spectrum and identifying the protein were submitted to MS/MS analysis on the Ultraflex MALDI-TOF/TOF by using LIFT™ technology (Bruker Daltonics).

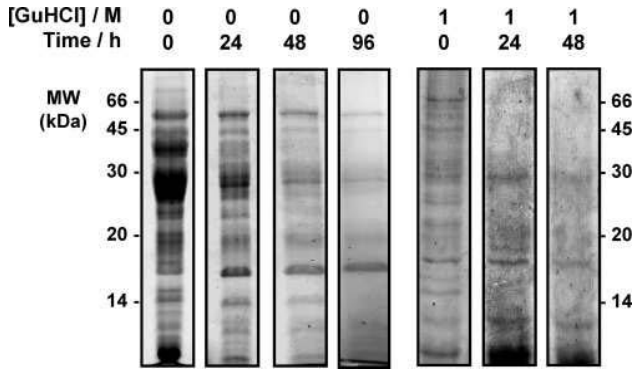


Fig.3.2 SDS-PAGE analysis of the *Sulfurisphaera* cytosolic proteome. The SulfCP72 cytosolic extract was incubated at 90 °C up to 96 h either in the absence or in the presence of GuHCl. Each lane was loaded with ≈1 µg of protein. Gels were Silver-stained.

Miscellaneous biochemical and spectroscopic methods

Protein purity was monitored by 12.5% SDS-PAGE. Gels were stained with Coomassie brilliant blue G or silver (Silver staining kit-protein, Amersham Biosciences) and images were obtained on ImageScanner (Amersham Biosciences). Protein content was determined by Bradford (24) quantification method. UV and visible spectra were recorded on UV-1700 Shimadzu spectrophotometer. Folded structure of ferredoxin purified from the thermally treated proteome was tested with fluorescence spectroscopy measurements, (Cary Varian Eclipse spectrofluorimeter) monitoring tryptophan fluorescence emission between 300-450 nm, with excitation at 280 nm. Peroxidase activity of the Bacterioferritin co-migratory protein (BCP) was tested by its ability to remove added H_2O_2 from a reaction mixture containing 50 mM HEPES pH 7 and 10 mM DTT as electron donor, at 25°C using an amperometric H_2O_2 detection (WPI Apollo 4000 system). Assay of SOD activity was done at 25ºC by the xanthine/xanthine oxidase method(25) in 50 mM phosphate buffer pH 7.8 with 0.1 mM EDTA; one activity unit was defined as the amount of enzyme that caused 50% inhibition of the cytochrome c reduction by superoxide (O_2^-).

3.4. Results and Discussion

High temperature and chemical denaturants induce proteome perturbation

The cytosolic proteomes obtained from *Sulfurisphaera* cells grown a 72°C (SulfCP72) and 92°C (SulfCP92) were perturbed by two different protocols. The thermal perturbation protocol consisted in incubation at 90°C for up to 96 h, sampling every 24 h. After 96 h incubation, the protein yield was 5%. In order to detail the enrichment of the thermostable proteome, a combination of small scale ionic exchange and SDS-PAGE was used. Aliquots of the SulfCP72 proteome, which had been sampled during different periods of the thermal perturbation protocol, were applied into a HiTrap Q-sepharose FF column for chromatographic separation. From the obtained protein chromatograms it became clear that the extract is being selectively enriched in thermostable proteins, which eluted at identical ionic strengths despite the longer incubation at 96°C (Fig. 3.1).

This provided a good indication that the tertiary fold of the proteins in solution was kept: in fact protein unfolding results in the exposure of a higher number of side chains thus changing the charged surface in respect to the native state, resulting in a modification of the elution profile which is not the case. The global impact of the thermal perturbation on the proteome was also investigated by SDS-PAGE analysis (Fig. 3.2). An increasingly lower number of bands were observed as the incubation time increases, evidencing protein selection. Identical results were obtained for cells grown at 92°C. The second perturbation protocol involved the combined use of high temperature (90°C) and a moderate concentration (1 M) of the chemical denaturant guanidinium hydrochloride (GuHCl). SDS-PAGE analysis shows that the presence of denaturant is sufficient to significantly reduce the time necessary to produce a subset of the proteome (24 and 48 h) (Fig. 3.2). Interestingly, a comparable profile was obtained from both perturbation protocols on SDS-PAGE, suggesting a common selection towards proteins with enhanced stability properties.

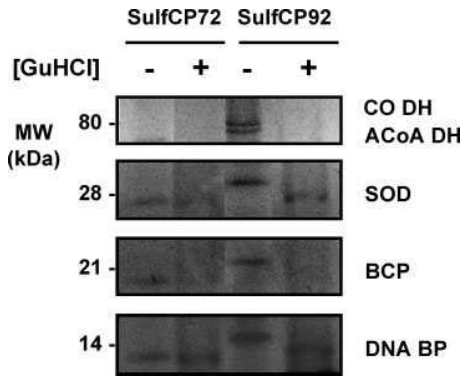


Fig.3.3 Detail of a 13 cm resolving 12.5% SDS-PAGE Silver-stained gel highlighting proteins whose excision led to their identification by mass spectrometry. The SulfCP72 and SulfCP92 cytosolic proteomes were incubated during 24 h in the presence (+) or absence (-) of 1 M GuHCl. The chemical denaturant was removed by extensive dialysis prior to electrophoresis. DNA BP, DNA binding protein Alba-2; BCP, bacterioferritin comigratory protein; SOD, superoxide dismutase; CO DH, hypothetical CO dehydrogenase large subunit; and ACoA DH, hypothetical acyl-CoA dehydrogenase. See also Table 1 for details.

In order to further investigate and identify some of these proteins, 13cm gels were prepared from SulfCP72 and SulfCP92 which had been incubated for 24h at 90ºC in presence of 1M GuHCl as well as in its absence (Fig. 3.3). From these gels, relevant bands were excised and subjected to MALDI-TOF/TOF analysis, leading to the identification of five proteins: a DNA binding protein 7e (DNA-BP, ≈14kDa), the bacterioferritin co-migratory protein (BCP, ≈21kDa), a superoxide dismutase (SOD, ≈28kDa), the large subunit of an hypothetical CO dehydrogenase (CO DH, ≈80kDa) and a hypothetical Acyl-CoA dehydrogenase (ACoA DH, ≈78kDa) (Fig. 3.3 and Table 3.1). These proteins have enhanced stabilities in respect to the total proteome, but they are not all equally stable: for example, whereas the DNA binding protein and the SOD are found in thermally and chemically perturbed extracts, the BCP and the dehydrogenases are more susceptible to simultaneous chemical degradation. In fact, the slight differences

observed in the electrophoretic mobility of these proteins, may be suggestive of moderately perturbed conformational states.

Proteome analysis by 2-DE and MS: identification of hyperstable proteins

The cytosolic fractions perturbed during 96h at 90°C were further resolved by two-dimensional electrophoresis. The SulfCP92 proteome was analyzed before and after the thermal perturbation (Fig. 3.4): it became clear that the number of proteins decreases substantially and that even after such extensive incubation there is still a large number of proteins detected. Proteins present in the native (Fig. 3.4A) and thermally perturbed (Fig. 3.4B) SulfCP92 proteome were selected for further investigation by MALDI peptide mass fingerprinting. Excised gel spots were digested and submitted to MALDI peptide mass fingerprinting analysis.

spot	protein annotation	MW (kDa)	pI	homologue accession Number
	Cellular Processes/Detoxification			
1	Superoxide dismutase *(a)*	24.3	3.2	15922615
2	Bacterioferritin comigratory protein (Peroxiredoxin)	17.7	7.5	15922097
	DNA Binding/Translation/Protein Modification			
3	DNA binding protein Alba-2*a*	10.5	n.d.	68567728
4	hypothetical DNA-binding protein 7e	7.3	10	15621644
5	hypothetical elongation factor 1-alpha	48.3	10	15920458
6	Thermosome subunit	60.7	5.4	15920519
	Energy Metabolism			
7	hypothetical CO dehydrogenase large subunit*(a)*	81.8	n.d.	15922093
8	hypothetical CO dehydrogenase middle subunit	31.2	10	15922095
9	hypothetical acetil-CoA synthetase	74.5	n.d.	BAB65737
10	hypothetical glutamate dehydrogenase (hyp)	45.7	6.1	15922573
11	hypothetical sulphide dehydrogenase	43.0	6.8	15920835
12	hypothetical L-lactate dehydrogenase	34.7	6.3	15622912
13	Ferredoxin [3Fe4S][4Fe4S]	11.3	5.5	2554684
14	hypothetical metal-dependent hydrolase	25.5	5.8	22096255

(a) Denotes proteins identified also from 1D SDS-PAGE gel bands.
n.d., not determined.

Table 3.1 Selected proteins identified from the cytosolic *Sulfurisphaera* proteome after perturbation

In order to confirm protein identifications, the lists of the identifying peptide masses deriving from each spot were carefully inspected searching for the best s/n peptides signal to sequence. The structural

information deriving from both the peptide mass fingerprinting and the MS/MS derived peptide sequences greatly enhanced the protein identification confidence. The subset of proteins identified from analysis of the 2DE gel (Table 3.1) can be grouped into three distinct functional categories: proteins involved in cellular processes and detoxification, DNA binding, translation and protein modification and energy metabolism.

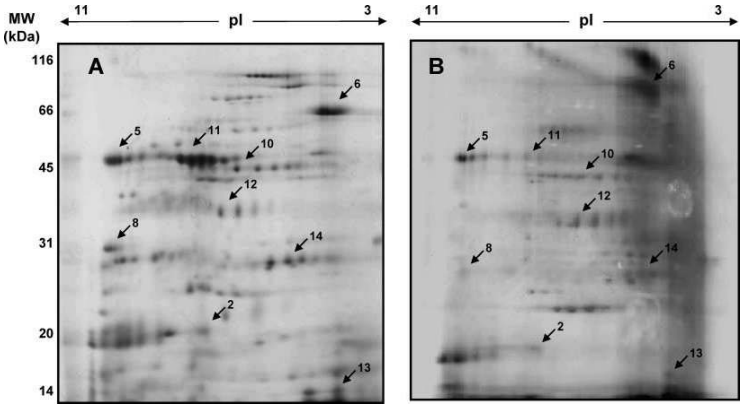


Fig. 3.4. 2DE Silver-stained gel (12.5% SDS-PAGE) using a 13 cm, 3-11NL IPG strip, resulting from the application of 500 μg of total protein. (A) Native SulfCP92 without any heat treatment. (B) Subset of the perturbed proteome obtained after 96 h incubation at 90 °C. Numbered arrows indicate spots which were selected for identification (see Table 3.1 for details).

Two of the identified proteins, superoxide dismutase and bacterioferritin co-migratory protein (BCP, a peroxiredoxin) had already been identified from the 1D gel analysis. Four identified proteins (DNA binding protein Alba-2, hypothetical DNA-binding protein 7e, hypothetical elongation factor 1-alpha, Thermosome β subunit) are involved in nucleic acid processing and protein modification processes. In particular the DNA binding protein Alba-2 was previously identified by analysing 1D gel of perturbed proteome. Finally, eight proteins involved in energy metabolism processes were identified, among which are several hypothetical dehydrogenases and the iron-sulfur protein ferredoxin (Table 3.1). MALDI MS and MS/MS spectra of superoxide dismutase, DNA binding protein Alba-2 and ferredoxin, each belonging to one of the three

different biological classes have been selected as an example and reported in Fig. 3.5.

The type of cellular processes in which the identified proteins are involved is very interesting and noteworthy from the point of view of the *Sulfurisphaera's* optimal growth conditions at high temperatures. The identified proteins participate in cellular processes (e.g. defense against reactive oxygen species, nucleic acid protection and energy production) in which some key proteins have enhanced thermal stabilities, which may relate to their importance on the metabolism of a thermophile. For example, at temperatures above 70^{o}C, DNA is particularly susceptible to chemical modifications, mainly depurination followed by cleavage of the nearby phosphodiesteric bond, thus making DNA binding proteins particularly important as these wind and compact DNA, protecting it (1). In agreement, several DNA-binding proteins were identified in this study (Table 3.1). Concerning the energy metabolism proteins, the hypothetical dehydrogenases identified suggest that some electron transfer pathways involved in energy production processes involve particularly stable proteins.

A common feature among these proteins is that they comprise a NAD(P)-binding Rossman fold: however, since this is one of the most populated folds in proteins, typical among dehydrogenases, a direct correlation with enhanced protein stability can not be postulated. Finding proteins involved in cellular detoxification processes to be hyperstable may be rationalized with the fact that high temperatures and aerobic growth conditions are prone to an increased cellular oxidative stress. In these circumstances, the amount of reactive oxygen species (ROS) formed during electron transfer processes is likely to be increased as a result of electron leakage from redox proteins and reduced cofactors. This calls for a particularly stable set of ROS defense proteins such as peroxiredoxin and superoxide dismutase (Table 3.1). Altogether, these results suggest that some metabolic processes in thermophiles may require a 'thermostable character' in order to cope with the specificities of life at high temperatures.

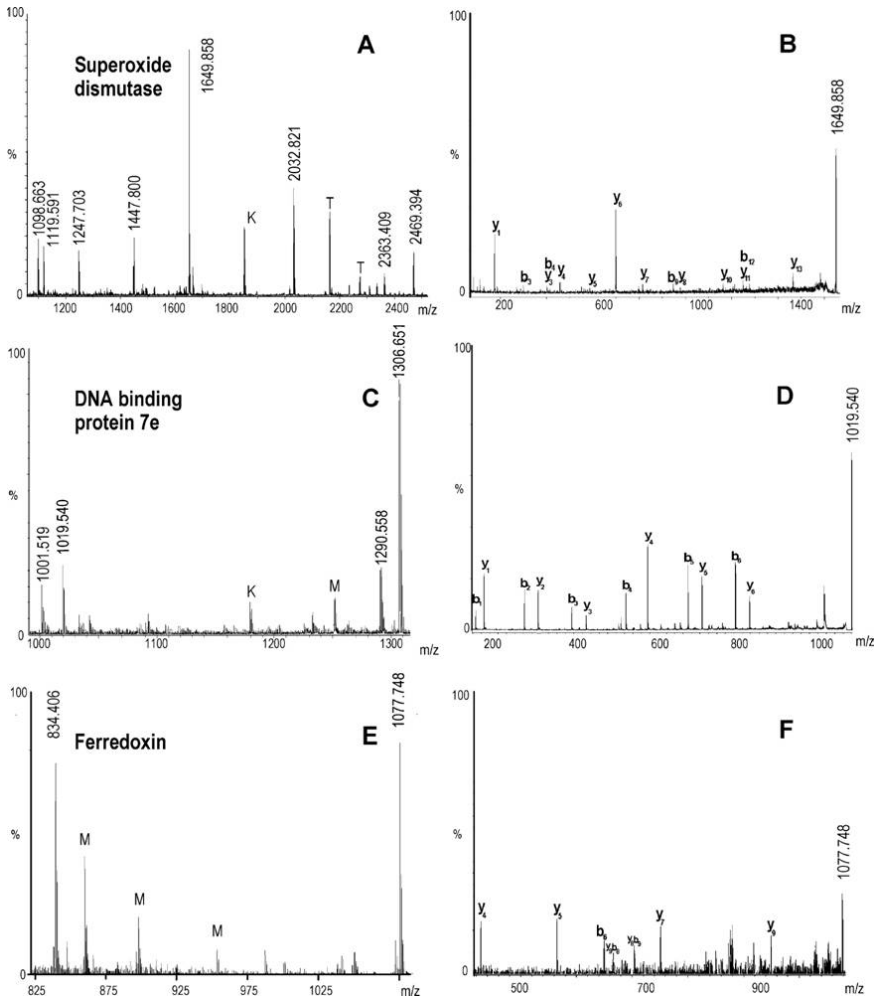


Fig. 3.5. MALDI MS and MALDI MS/MS spectra of 3 out of 14 identified proteins from *Sulfurisphaera*. (A, C, and E) Panels show the MALDI MS spectra of superoxide dismutase, hypothetical DNA-binding protein, and ferredoxin, respectively. K, M, and T stand for keratin, matrix, and trypsin mass signals, respectively. (B, D, and F) Panels show the corresponding MALDI MS/MS spectra of the above proteins where a peptide mass signal was chosen for each protein (1649.858, 1019.540, and 1077.748, respectively) and submitted to tandem MS in order to confirm protein identifications.

Proteins from the pool of selected hyperstable proteins are biologically active

Thermophilic proteins can be intrinsically very stable and exhibit a very high kinetic stability towards thermal degradation. We have validated our proteome perturbation method by verifying the extent to which the subset of proteins obtained after the perturbation corresponds to biologically active molecules.

Three distinct proteins were selected from the perturbed subset: the iron-sulfur protein ferredoxin (Fd), the peroxiredoxin Bacterioferritin co-migratory protein (Prx-BCP) and superoxide dismutase (SOD). Whereas Fd is an electron carrier present in most thermoacidophilic Archaea, Prx-BCP and SOD are enzymes that are respectively involved in the detoxification of peroxides and superoxide. For this characterization we have set up a small-scale two-step chromatographic procedure for partly resolving the thermally perturbed SulfCP92 proteome. Starting from the 96h incubated SulfCP92 proteome, and using a combination of anionic exchange (Q-sepharose HiPrep) and gel filtration (S-75) a protein fraction corresponding to a molecular weight of ≈12 kDa, as determined from SDS-PAGE, was obtained. Sequencing of its N-terminus (1-GIDPNYRTNRQVVGEHSK-G) clearly showed that this protein corresponds to ferredoxin. A preliminary biophysical characterization was carried out on this purified band, and its assignment as a Fd was corroborated by obtaining the typical UV-visible spectrum, with the characteristic band at 410 nm corresponding to native Fd with intact Fe-S clusters (not shown).

The *Sulfurisphaera* ferredoxin has orthologues in other thermophilic *Sulfolobales* (11, 16) and these proteins are characterized by a very high thermal stability: they have very high midpoint melting transitions (T_m=110°C at pH 7 for the *Acidianus ambivalens* Fd) as well as a very high kinetic stability(11). The protein core harboring the two iron-sulfur centers is likely to play an important role in protein thermostability, as well as a Zn-containing N-terminal extension(11, 26).

The SOD (27) and Prx-BCP (28) activities were measured in the native and perturbed SulfCP72 proteome. Not only both activities were present but also a specific activity increase of the enzymes was observed (Fig. 3.6).

This observation is compatible with a significant enrichment of these two enzymes in the thermally treated extract, as a result of their enhanced thermostability. Whereas little is yet known concerning the molecular origins of Prx-BCP stability, there are a several studies on thermophilic SODs(27). The available data on Archaeal Fe-SODs corroborates our finding that these proteins are among the most thermostable in the cell. For example, the S. acidocaldarius SOD, which is a functional tetramer, has a very high stability undergoing structural thermal melting at 125°C, a temperature which is 40°C above the organism optimal growth temperature(27). Structural analysis suggests that the increase protein stability may result from its quaternary structure, a tight packing of buried hydrophobic residues and an increased number of ion pairs (27).

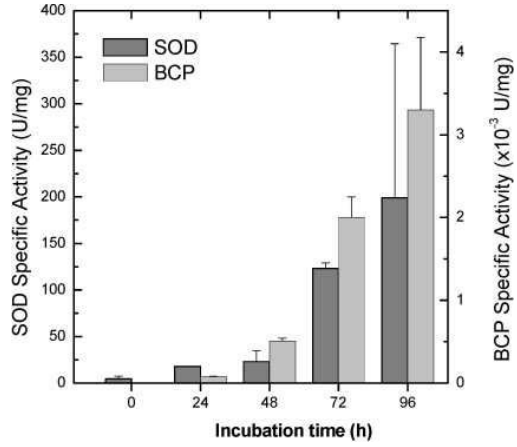


Fig. 3.6. Variation of the BCP and SOD specific activity as a function of incubation time of the SulfCP92 proteome. Error bars are from the standard deviation of activity assays (n =2).

3.5. Conclusions

Here we have outlined an experimental strategy which allows mining proteomes for proteins having enhanced stability properties. This approach is particularly valuable for the identification of proteins which can be used in subsequent studies aimed at the understanding of protein stability, function or structural features. Following our interest in the study of conformational properties of thermophilic proteins, we have selected *Sulfurisphaera sp.* as a model organism to implement this

methodology: the identification of hyperstable proteins in a thermophilic background required drastic perturbation protocols (e.g. 4 days at 90°C) which nevertheless resulted in the identification of a subset of proteins which remained folded after the perturbation and provided an insightful perspective into the type of essential cellular processes requiring particularly resistant proteins. In fact, many of the identified proteins are involved in stress response mechanisms that aim at protecting nucleic acids and proteins from aggression elements such as thermal stress and reactive oxygen species. In this respect, this approach highlighted not only proteins interesting for subsequent stability studies, but also on key metabolic processes which may have themselves a 'thermophilic character'. Recently, efficient expression systems have been developed (29) using the closely related hyperthermophile *Sulfolobus solfataricus* as an efficient expression system of tagged proteins, thus greatly facilitating the production of proteins of interest such as those identified in this work for subsequent studies.

3.6. References

1. **Madigan, M. T., and A. Oren.** 1999. Thermophilic and halophilic extremophiles. Curr Opin Microbiol **2:**265-9.
2. **Vieille, C., and G. J. Zeikus.** 2001. Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev **65:**1-43.
3. **Martins, L. O., R. Huber, H. Huber, K. O. Stetter, M. S. Da Costa, and H. Santos.** 1997. Organic Solutes in Hyperthermophilic Archaea. Appl Environ Microbiol **63:**896-902.
4. **Petsko, G., and D. Ringe.** 2004. Protein Structure and Function. New Science Press.
5. **Cowan, D. A.** 1992. Biotechnology of the Archaea. Trends Biotechnol **10:**315-23.
6. **Li, W. F., X. X. Zhou, and P. Lu.** 2005. Structural features of thermozymes. Biotechnol Adv **23:**271-81.
7. **Fitter, J.** 2005. Structural and dynamical features contributing to thermostability in alpha-amylases. Cell Mol Life Sci **62:**1925-37.
8. **Gomes, C. M., A. Kletzin, and M. Teixeira.** 2002. An archaeal b-type cytochrome containing a nonfunctional carbonic anhydrase-like domain. J Biol Inorg Chem **7:**483-9.
9. **Perl, D., and F. X. Schmid.** 2001. Electrostatic stabilization of a thermophilic cold shock protein. J Mol Biol **313:**343-57.

10. **Urich, T., C. M. Gomes, A. Kletzin, and C. Frazao.** 2006. X-ray Structure of a self-compartmentalizing sulfur cycle metalloenzyme. Science **311:**996-1000.
11. **Gomes, C., A. Faria, J. Carita, J. Mendes, M. Regalla, P. Chicau, H. Huber, K. Stetter, and T. M.** 1998. Di-cluster, seven-iron ferredoxins from hyperthermophilic Sulfolobales. JBIC **3:**499-507.
12. **Gomes, C. M., C. Frazao, A. V. Xavier, J. Legall, and M. Teixeira.** 2002. Functional control of the binuclear metal site in the metallo-beta-lactamase-like fold by subtle amino acid replacements. Protein Sci **11:**707-12.
13. **Gomes, C. M., J. B. Vicente, A. Wasserfallen, and M. Teixeira.** 2000. Spectroscopic studies and characterization of a novel electron-transfer chain from Escherichia coli involving a flavorubredoxin and its flavoprotein reductase partner. Biochemistry **39:**16230-16237.
14. **Henriques, B. J., L. M. Saraiva, and C. M. Gomes.** 2005. Probing the mechanism of rubredoxin thermal unfolding in the absence of salt bridges by temperature jump experiments. Biochem Biophys Res Commun **333:**839-44.
15. **Henriques, B. J., L. M. Saraiva, and C. M. Gomes.** 2006. Combined spectroscopic and calorimetric characterisation of rubredoxin reversible thermal transition. J Biol Inorg Chem **11:**73-81.
16. **Leal, S. S., and C. M. Gomes.** 2005. Linear three-iron centres are unlikely cluster degradation intermediates during unfolding of iron-sulfur proteins. Biol Chem **386:**1295-300.
17. **Leal, S. S., M. Teixeira, and C. M. Gomes.** 2004. Studies on the degradation pathway of iron-sulfur centers during unfolding of a hyperstable ferredoxin: cluster dissociation, iron release and protein stability. J Biol Inorg Chem **9:**987-96.
18. **Wittung-Stafshede, P., C. M. Gomes, and M. Teixeira.** 2000. Stability and folding of the ferredoxin from the hyperthermophilic archaeon Acidianus ambivalens. J Inorg Biochem **78:**35-41.
19. **Kurosawa, N., Y. H. Itoh, T. Iwai, A. Sugai, I. Uda, N. Kimura, T. Horiuchi, and T. Itoh.** 1998. Sulfurisphaera ohwakuensis gen. nov., sp. nov., a novel extremely thermophilic acidophile of the order Sulfolobales. Int J Syst Bacteriol **48 Pt 2:**451-6.
20. **Kawarabayasi, Y., Y. Hino, H. Horikawa, K. Jin-no, M. Takahashi, M. Sekine, S. Baba, A. Ankai, H. Kosugi, A. Hosoyama, S. Fukui, Y. Nagai, K. Nishijima, R. Otsuka, H. Nakazawa, M. Takamiya, Y.**

Kato, T. Yoshizawa, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki, S. Masuda, M. Yanagii, M. Nishimura, A. Yamagishi, T. Oshima, and H. Kikuchi. 2001. Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, Sulfolobus tokodaii strain7. DNA Res **8:**123-40.

21. **Chong, P. K., and P. C. Wright.** 2005. Identification and characterization of the Sulfolobus solfataricus P2 proteome. J Proteome Res **4:**1789-98.
22. **Urich, T.** 2001. Ein neuer Stoffwechselweg bei Archaeen: Untersuchungen zur anaeroben Schwefeldisproportionierung bei Acidianus ambivalens und Sulfurisphaera sp.. , . Diploma Thesis. Darmstadt University of Technology, Darmstadt
23. **Teixeira, M., R. Batista, A. P. Campos, C. Gomes, J. Mendes, I. Pacheco, S. Anemuller, and W. R. Hagen.** 1995. A seven-iron ferredoxin from the thermoacidophilic archaeon Desulfurolobus ambivalens. Eur J Biochem **227:**322-7.
24. **Bradford, M. M.** 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem **72:**248-54.
25. **McCord, J. M., and I. Fridovich.** 1969. Superoxide dismutase. An enzymic function for erythrocuprein (hemocuprein). J Biol Chem **244:**6049-55.
26. **Rocha, R., S. Leal, V. Teixeira, M. Regalla, H. Huber, A. Baptista, C. M. Soares, and C. M. Gomes.** 2006. Natural domain design: enhanced thermal stability of a zinc lacking ferredoxin isoform shows that a hydrophobic core efficiently replaces the structural metal site. Biochemistry ***in press***.
27. **Schafer, G., and S. Kardinahl.** 2003. Iron superoxide dismutases: structure and function of an archaic enzyme. Biochem Soc Trans **31:**1330-4.
28. **Limauro, D., E. Pedone, L. Pirone, and S. Bartolucci.** 2006. Identification and characterization of 1-Cys peroxiredoxin from Sulfolobus solfataricus and its involvement in the response to oxidative stress. Febs J **273:**721-31.
29. **Albers, S. V., M. Jonuscheit, S. Dinkelaker, T. Urich, A. Kletzin, R. Tampe, A. J. Driessen, and C. Schleper.** 2006. Production of recombinant and tagged proteins in the hyperthermophilic archaeon Sulfolobus solfataricus. Appl Environ Microbiol **72:**102-11.

The results presented in this chapter resulted of collaboration with following laboratories and people:

Collaboration with Environmental Systems Group, Department of Chemical and Process Engineering, University of Sheffield, UK, headed by Prof Phillip Wright. iTRAQ experiments were performed in collaboration with Dr Trong Khoa Pham during a visit period in November-December 2008.

Collaboration with Computational Genomics Laboratory, Instituto Gulbenkian de Ciência, Oeiras, Portugal headed by Dr José B. Perreira Leal. Member of this group, Renato Alves, provided computational algorithm for data processing.

Chapter 4

INTRINSIC THERMAL STABILITY PROPERTIES IN THERMOPHILIC VS. MESOPHILIC CYTOSOLIC PROTEOME: THERMAL SEPARATION, IDENTIFICATION AND iTRAQ QUANTIFICATION

Intrinsic thermal stability properties in thermophilic vs mesophilic cytosolic proteome: Thermal separation, identification and iTRAQ quantification

4.1. Summary

The comparative proteomic analysis of extremely thermophilic organism vs. mesophilic one offers the opportunity to discover strategies for maintaining genome and proteome integrity under the harsh conditions, shedding light on the special adaptations that are essential for life at high temperature. Therefore, an extremophile *Sulfolobus solfataricus* and a mesophile *Escherichia coli* were chosen as model systems for this type of analysis that included exhaustive thermal treatment of soluble cytosolic proteome in order to isolate a highly stable proteome subset. Soluble cytosolic extract of these organisms were exposed to high temperature up to 90^{o}C during prolonged period of time and proteins that remained soluble after this treatment were subjected to further analysis. iTRAQ protein sequences identification and quantification of this "thermoproteome" and bioinformatics analysis gave an insight into possible relationship between thermostability on one side and amino acid content, physicochemical properties and biological function on the other. A general observation regarding these results is that none of the so-called predictive thermostability measures have strong predictive capabilities and that most obvious explanation is that none of them *per se* are the major factor contributing protein stability but possibly combination of these factors. Further analysis of the proteomic level of the defined thermally enhanced groups is advised, including these as well as the other thermophilic and mesophilic organisms, in order to asses the proteomic basis of thermostability in more detail. Evolutionary relationships between the groups as well as protein features can further be discussed from the broader environmental point of view in order to assess the ecological and evolutionary aspect of thermostability.

4.2. Introduction

This work was focused on identifying the basis of protein stability to high temperatures as integral to our understandings of protein folding, the relationship of protein structure to function, and the evolution of life on this planet. Most published studies of overall protein thermostability take one of two alternative approaches. The first approach is structural/mutational: it uses protein structures in order to find differences between thermophilic and mesophilic proteins and thereby to propose hypotheses for the bases of thermal adaptation. In continuance, many of these hypotheses further guide directed mutagenesis studies that may yield a detailed insight of the interactions that are stabilizing a particular protein. Unfortunately, the work that involves such structural/mutational approach often restricts analyses to more or less limited numbers of proteins, and offers a biased view of overall thermal stabilizing mechanisms. In addition, the lack of consensus among studies has given rise to the recognition that no set of simple factors distinguish between thermophile and mesophile proteins but this rather involves multiple interaction of various determinants. In order to address the determinants of thermostability a broader approach to the problem is required. The second, more comprehensive approach, is based on the comparisons of families of proteins isolated from thermophilic and mesophilic organisms with consequent statistical analysis of the achieved results. In such comparative studies, protein features are discussed from the broader environmental point of view that distinguish proteins that *in vivo* optimally exist and function, possibly assessing the ecological and evolutionary aspect.

Analysis of extremely thermophilic archaea offers the opportunity to discover strategies for maintaining genome and proteome integrity of the relatively little explored third domain of life, thereby shedding light on the diversity and evolution of these central and important systems. These studies also reveal special adaptations that are essential for life at high temperature. A number of investigations of the hyperthermophilic and acidophilic crenarchaeota *Sulfolobus solfataricus* have been performed over the years. The property of *Sulfolobus solfataricus* is to grow in extreme conditions like high temperature of 80°C and in highly

acidic environments of pH 2−4. Impressive ability to survive and thrive in extreme environments and therefore its proteome represents a good working model for life under extreme conditions. It has a genome size of ~ 3.0 Mbps and approximately 3000 predicted open reading frames (ORFs) (1). The *S. solfataricus* proteome has already been studied using a variety of techniques such as two-dimensional gel electrophoresis (2) and shotgun proteomics, in order to give information regarding cell cycle, transcription, RNA processing and translation, DNA replication, and several metabolic pathways.

A mesophilic bacterium *Escherichia coli* has been considered a model organism and has historically been the focus of many studies. Many decades of research have resulted in a wealth of genetic, biochemical, and structural information that together can be paralleled in other systems. Due to its availability it has an extraordinary position as a preferred model in biochemical genetics, molecular biology, and biotechnology and was the earliest organism to be suggested as a candidate for whole genome sequencing (3). It was and is the primary model organism for bacteria, used to define the genetic code, elucidate various cellular mechanisms of central importance such as transcription, translation, restriction, replication, and much of basic metabolism, being often both a tool and object of study of genome science. *E. coli* is a facultative anaerobe and a metabolic opportunist, spending part of its natural life cycle living anaerobically in the intestinal tracts of animals, part living aerobically in diverse natural environments such as rivers or soil, and, in the case of pathogens, invading animal or human hosts. Individual strains of *E. coli* vary in their preferences for niches/hosts, and these variations are reflected in their gene contents. It has been estimated that the genomes of natural isolates of *E. coli* range from 4.5 to 5.5 Mbps (4-6). It has been somehow an obvious choice for us to include such well studied organism as the reference in order to compare the information gained in the study of *S. solfataricus*.

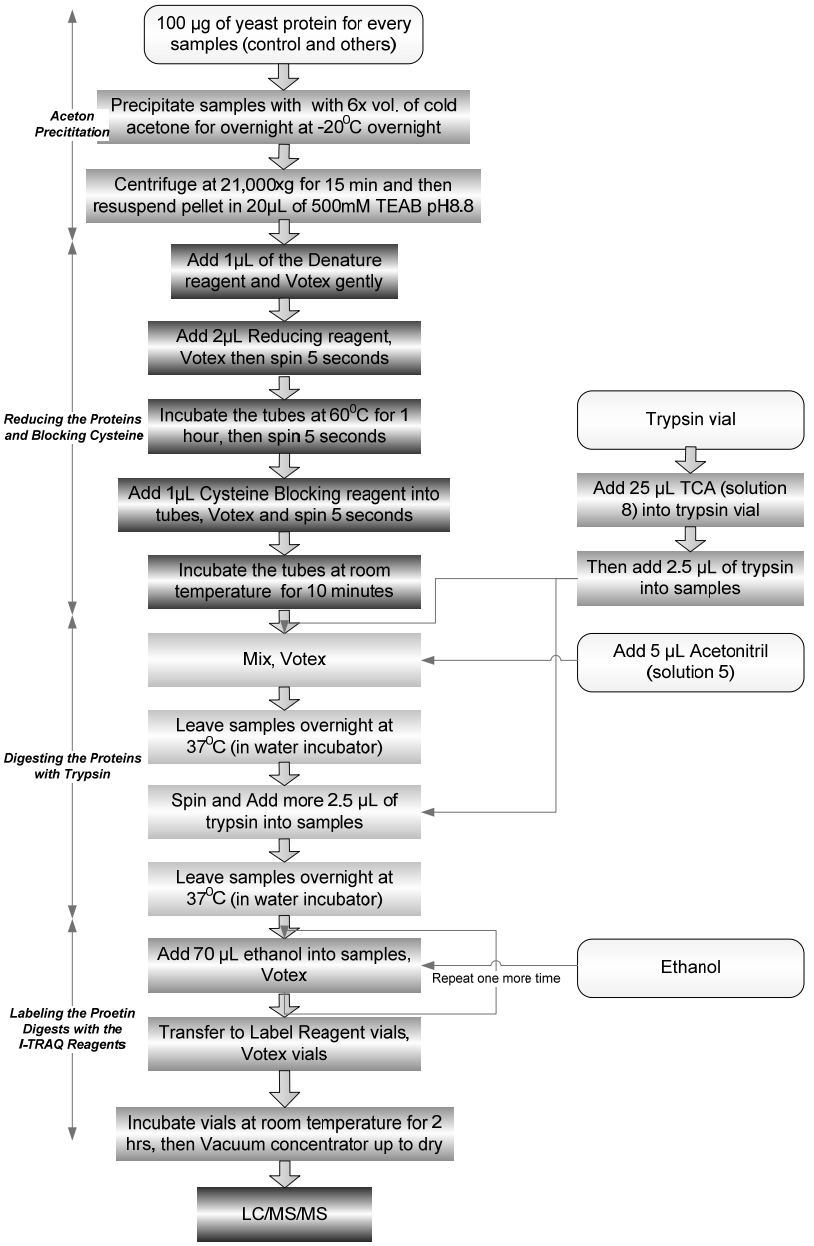
4.3. Materials and methods

The soluble extracts from *S. solfataricus* and *E. coli* were perturbed by two different protocols in 40 mM phosphate buffer, pH 6.5. The thermal

perturbation of *S. solfataricus* included incubation of soluble extract at 90°C for up to 96 h and sampling every 24 h. In the perturbation protocol of *E. coli*, soluble extract was incubated also at 90°C and sampled every 5 min up to 1 hour of incubation. In both cases, after sampling, less stable precipitated proteins were removed by 10 min centrifugation at 14 000 rpm on a bench centrifuge at 4°C. Thermally isolated remaining soluble protein sequences were further identified and subjected to iTRAQ quantification, followed by bioinformatics data analysis.

The chosen methodology for our study included exhaustive thermal treatment of soluble cellular extract followed by identification and iTRAQ quantification of remaining soluble proteins and further data analysis. iTRAQ - isobaric tags for relative and absolute quantification method, is a successful and widely used technique for identification and quantification (7).The iTRAQ reagent consists of a reporter group, a balance group and a peptide reactive group. The reactive group specifically reacts with the N-terminus and side-chain amines of peptides. The reporter group is a tag with various masses. Four different masses allow 4-plex experiments, based on various combinations of isotopic elements (Scheme 4.1). Recently 8-plex implementations have become commercially available (8, 9). The balance group varies also in mass to ensure that the combined mass of the reporter and balance group remains constant. As such, peptides labeled with different isotopes are isobaric in MS, while the reporter groups are released during CID fragmentation, generating a low molecular mass reporter ion used for the relative quantification of the specific peptides and proteins from which it originates. iTRAQ reagents with different tags are used to label peptides either from different biological conditions, or in our case differently thermally treated sample.

In comparison to 2DE this approach offers significant reductions in labor intensity and the overall time taken to analyze the proteome and also demonstrates this methodology as an alternative approach for proteomic studies or in combination with 2DE. iTRAQ analysis was performed in collaboration with Environmental Systems Group, Department of Chemical and Process Engineering, University of Sheffield, UK.



Scheme. 4.1 iTRAQ protocol, courtesy of Dr. Trong Khoa Pham, University of Sheffield

In our study, after iTRAQ quantification of thermally selected sequences, they were compared from various points of view: molecular weight, isoelectic point, amino acid content, protein class and COG functional categories, freely available on the internet, in order to assess various aspects of elevated thermostability of these sequences. Retrieving these parameters from the internet was facilitated by using computational algorithms through Computational Genomics Laboratory, Instituto Gulbenkian de Ciência, Oeiras, Portugal.

Defining thermostable proteins' subsets

Thermally isolated soluble protein sequences were selected by incubation at 90^{o}C, further subjected to identification and iTRAQ quantification, followed by bioinformatics data analysis. Reasonable selection of subgroups for the further analysis was the initial issue. Therefore, we have defined the isolated and identified protein subsets as follows:

"*Survivors*"- Full set of iTRAQ quantifie proteins at the end of the 90 C thermal treatment for both organisms, consisting of 326 identified sequences for *S. solfataricus* and 364 for *E. coli*. Therefore, "*Survivors*", and indeed they are, were the **all** of the sequences from the soluble proteome that remained soluble after the applied thermal treatment at 90^{o}C (Fig. 4.1).

"*Unchanged*"- subgroup of the "*Survivors*", are isolated and identified sequences with almost unchanged relative content in respect to the time 0 of the incubation period. Relative enhancement ratio between 0.6 and 1.5 was decided to delineate the boundaries of this subgroup. "*Unchanged*" group was represented with 105 identified *S. solfataricus* sequences and 92 *E. coli* sequences.

"*Thermally enhanced*" – subgroup of the "*Survivors*", are the identified sequences with the amazing increase in relative content between 1.5 and 4.35 times for *E. coli ,* and in the case of *S. solfataricus* relative increase was up to amazing 23.16 times. Sequences data are available in Appendix I and Appendix II for *E. coli* and *S. solfataricus* respectively.

Definition of the boundaries was based upon the decision and may

therefore be somewhat subjective. Illustrative demonstration of the isolated and iTRAQ identified subthermoproteome is represented in Fig. 4.2 for *E. coli*, and Fig. 4.3 for *S. solfataricus*.

E. coli		
Ec survivors	Ec unchanged	Ec thermally enhanced
n = 364	n = 92	n = 80
S. solfataricus		
Ss survivors	Ss unchanged	Ss thermally enhanced
n = 326	n = 105	n = 71
Definitions		
All proteins that were identified after thermal perturbation	Subset of the survivors whose relative proportion remains essentially invariable	Subset of the survivors relatively increased at the end of the thermal perturbation

Fig. 4.1 iTRAQ identified sequenses of the soluble cellular subproteome after thermal incubation at 90^0C. "Surviors" are all of the identified sequences of the soluble thermoproteome of respective organism. "Unchanged" are sequences with relative enhancement ratio between 0.6 and 1.5 in respect to the initial time of incubation. "Thermally enhanced" are sequences with increase in relative enhancement ratio more than 1.5 in respect to the time 0 of the incubation.

Identification and separation of these subcategories of thermally selected soluble proteome enabled us to take a better insight into obtained results. A criterion to separate the groups in this manner was subjective, and guided by amount of data provided by the experimental approach. Due to unexpectedly large number of identified sequences and amount of data, as well as the properties that we decided to compare and analyze, in this thesis are represented in more detail the properties of the sequences from the "*Thermally enhanced*" group of both investigated organisms, that are represented with 22% or roughly one quater of the initially identified protein sequences for both organisms respectively. Relative content of these sequences within the "*Thermally enhanced*" group has shown amazing increase up to 4.35 times for *E. coli,*

and in the case of *S. solfataricus* relative increase was up to amazing 23.16 times.

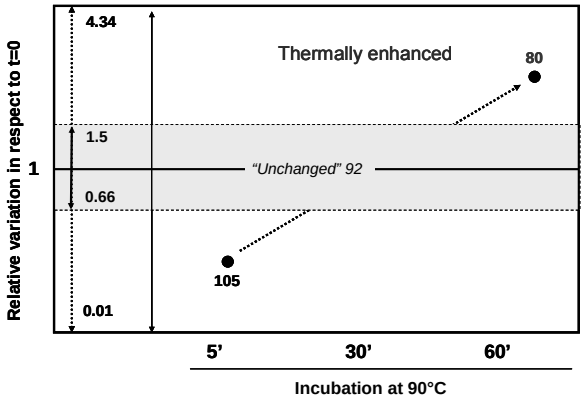


Fig. 4.2. Illustrative demonstration of the isolated and iTRAQ quantified thermoproteome for *E. coli,* 364 thermostable protein sequences identified. "*Thermally enhanced*" group is represented by 80 identified sequences.

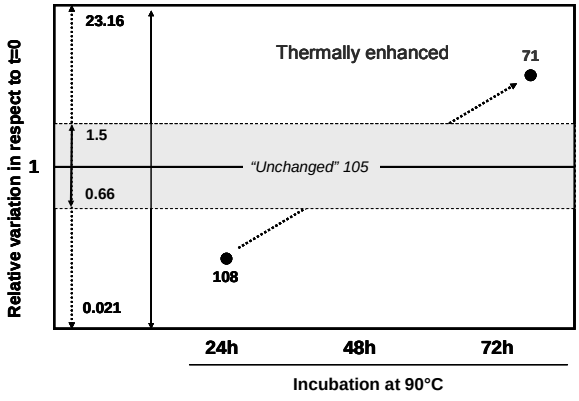


Fig. 4.3. Illustrative demonstration of the isolated and iTRAQ quantified thermoproteome for *S. solfataricus,* 326 thermostable protein sequences identified. "*Thermally enhanced*" group is represented by 71 identified sequences.

4.4. Relationship between thermostability and physicochemical properties

Physicochemical properties of isolated thermostable proteins, molecular weight (MW), isoelectric point (pI) and size, were addressed in order to access the possible correlation between these parameters and enhanced thermostabe properties.

It is important to keep in mind that considered MW and pI are theoretical according to amino acid sequence of the specific protein. In experimental gel based studies, estimation of MW and pI values from 2-D gels had been done mostly using a set of protein standards (10) or relative to a polymerized single standard (11). Although the reproducibility of 2-D gels is improved by IPG gels, this approach is still not sufficiently reliable which influenced our choice to deal with theoretical values determined from the specific sequence.

Molecular weight and isoelectric point

Regarding MW, there is no statistical support to any difference when compared to the full organism set of proteins. It can be noticed that for the both investigated organisms, identified sequences are rather small. Regarding isoelectric point (pI), *Thermally enhanced* protein subsets show a tendency towards bimodal distribution. It is in accordance with data from the literature that bimodal pI distribution in prokaryotic proteome (12) with peaks centered around pH 5.5 and pH 9 . This bimodality was explained as being caused by the fact that as proteins are least soluble at their pI, they have evolved to have pI's away from neutral pH – which was assumed to be the intracellular pH. Environmental preference of S*. solfataricus* is rather acidic habitats, and even though it's intracellular pH is physiological (pH=~7.6), environment might have an influence towards the selection of the highly stable proteome subset. Further in the literature, the presence of a trimodal pI distribution is observed in correlation of pI to intracellular localization: cytoplasmic, nuclear and membrane proteins seemed to lie largely in the acidic, neutral and basic portions of the trimodal distribution, respectively (13). Our data are in agreement with this for the both investigated organisms, as we are dealing with soluble cellular proteome subset, but it

unfortunately at this point still does not explain elevated thermostability of our isolated subproteomes.

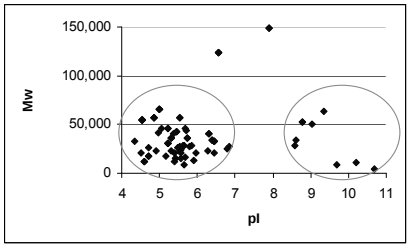


Fig. 4.4 Correlation between molecular weight (MW) and isoelectric point (pI) of Thermally enhanced subgroup of *E. coli*

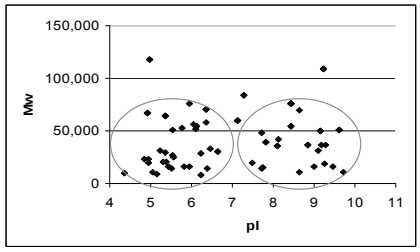


Fig. 4.5. Correlation between molecular weight (MW) and isoelectric point (pI) of Thermally enhanced subgroup of *S. solfataricus*

4.5. Relationship between thermostability and aminoacid content

All of our data regarding frequencies of certain aminoacids are compared with the same of the entire organisms' proteome in order to determine if the difference is statistically relevant. Therefore, in the figures in this chapter and corresponding figures, abbreviation "*R*" is used for the frequencies ratio of our results over the entire organism's proteome.

Aminoacid residue content was assessed in order to investigate the existence of prevalence of certain residues or classes within the most thermostable protein subsets (Figures 4.6 and 4.7) We were keeping in mind that data from the literature often claim that presence or absence of certain type of residues *per se* is not a sufficient prerequisite for difference in stability, but rather the position and subsequent interaction of the residues with the surrounding ones can lead towards alteration in stability properties.

For both investigated organisms, amino acid content was less frequent in Thermally enhanced proteome subgroup than in the full organisms' genome for the following amino acids: Phe, Leu, His, Tyr and Ile. On the other hand, more frequent were the following amino acids: Ala, Lys, Asp

and Glu.

Obvious was the tendency to include charged aminoacids : negatively charged (Asp and Glu), positively charged (Lys), but in contrast with His that are present in lower frequencies. Aromatic residues (Phe, Tyr) are in the lower proportions but not the same in the all of the datasets: Phe decrease is more pronounced at *S. solfataricus* while Trp decrease affects more *E coli* subsets. Observing alterations in aromatic and other residue content in our datasets, it can be only discussed that residues forming the additional cluster emanate from different secondary structural regions of the protein, stabilizing the tertiary fold. Regarding aromatic clusters, in studies comparing thermophilic and their mesophilic homologs, at least one additional aromatic cluster was found close to the active site of the thermophilic enzyme. The presence of additional aromatic clusters near the active site should help in retaining the conformational features of the active sites that is required to bind the substrate at high temperatures and thus contributing to the high thermophilicity of the thermostable proteins. Additional aromatic clusters occur in regular secondary structures, implying their location to be in more rigid regions of the protein (14). Obvious lower presence of these residues in our datasets may be explained by the fact that we are dealing here only with cytoplasmatic soluble fraction of the proteome, and that aromatic residues are found to be present with elevated incidence in membrane proteins (15) that we were not including in our study.

Cys in our datasets is increased with *S. solfataricus*, which is rather unusual. Frequency of Cys residues in thermophilic proteins in general is significantly decreased due to two reasons. Thermostability has, among other parameters, been attributed to enhanced secondary structure propensity. Cys is a helix disfavouring residue, and analysis of the composition of α-helices in the thermophilic proteins (16) have noted its significant decrease, together with His. The second reason is that Cys, together with Asn, Gln and Met can be classified as thermolabile due to their tendency to undergo deamidation or oxidation at high temperatures (17). So, disulfide bonds are not a method to achieve protein thermostability (18) and in general, hyperthermostable proteins

contain lower fractions of cysteines and histidines and are poorer in disulfide bonds than their thermostable and mesostable counterparts.

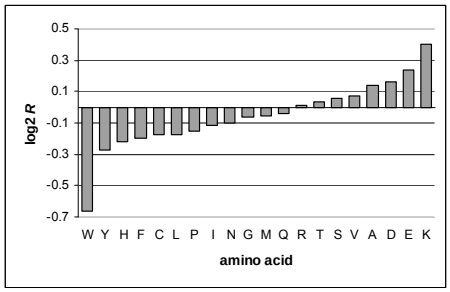


Fig 4.6. Thermally enhanced sequences subset of *E. coli*: Frequency of aminoacids in respect to the complete proteome of the non treated organism.

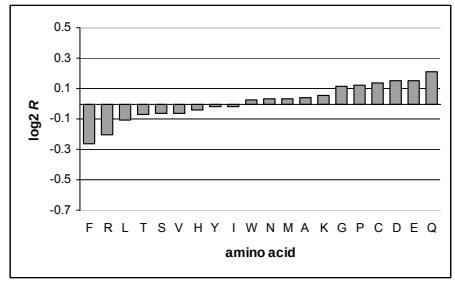


Fig 4.7. Thermally enhanced sequences subset of S*. solfataricus*: Frequency of aminoacids in respect to the complete proteome of the non treated organism.

All of this is supported by our data regarding *E. coli* Thermally enhanced subproteome, but interestingly our data are opposite from the literature for *S. sulfataricus*, probably because of the different mechanisms that enabled these sequences to achieve additional stability properties. The same observation applies to Pro, which is not consistent between the *S. solfataricus* and *E. coli* datasets, being more abundant in the first when compared to the full proteome of the organism, and less in the second.

Besides the decrease of Leu and Ile, other hydrophobic residues have not shown overall consistency through both datasets, even though it is generally accepted that hydrophobic effect is the main driving force in protein folding (19). Aliphatic amino acids contribute to the hydrophobic interaction, which is main force for maintaining conformational stability in inner part of protein (20).

It has been suggested that thermophilic proteins are substantially more hydrophobic and have more surface area buried upon oligomerization as compared with their mesophilic counterparts, even though overall hydrophobicity of thermophilic proteins and their mesophilic homologs are very similar (17). Our thermally isolated proteins are mostly small, as

opposite to large oligomeric proteins that would prefer more hydrophobic residues in the buried position that might somehow provide explanation to our results.

Besides from comparing the frequencies of individual amino acid residues, we have also compared and discussed frequencies of amino acid groups in according to their properties. Table 4.1 gives the list of groups. It was immediately obvious that polar and charged were more frequent in both organisms' Thermally enhanced datasets (Figures 4.8 and 4.9). It is in agreement with the data from other studies (17) that increased polar surface area and salt bridges contribute to the greater stability of the thermostable proteins.

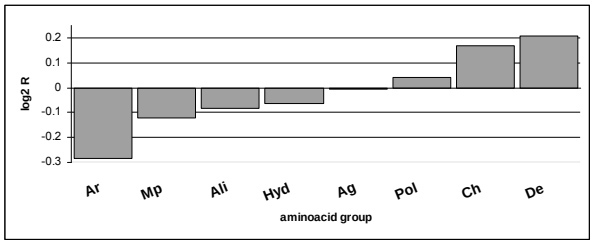


Fig 4.8 Thermally enhanced sequences subset of *E. coli:* Frequency of aminoacid groups in respect to the complete proteome of the non treated organism. Legend of the abbreviations is in Table 4.1.

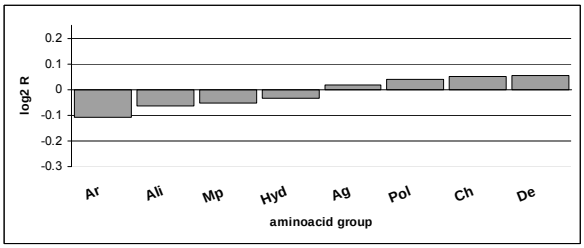


Fig 4.9 Thermally enhanced sequences subset of *S. solfataricus*: Frequency of aminoacid groups in respect to the complete proteome of the non treated organism. Legend of the abbreviations is in Table 4.1

Various published studies have shown the existence of the certain amino acids combinations as so-called “prediction factor” for thermostability. Farias and Bonato (21) published the theory that the ratio of E+K/Q+H was proved to be found in highly stable protein sequences. In most of our datasets this combination appears with significant difference to have predictive capacity (data not shown). On the other hand, combination of IVYWREL determined by Zeldovich et al (22) in our datasets was significantly decreased in five of six of of thermally enhanced datasets from both organisms which is an unexpected result (data not shown). This combination can even be used as a negative prediction factor, but this is in opposition of what has been expected.

legend	*properties*	*aminoacids*	
Ar	aromatic	Phe Trp Tyr	F W Y
Ali	aliphatic	Gly Ala Val Leu Ile	G A V L I
Hyd	hydrophobic	Gly Ala Val Ile Leu Met Phe Trp Cys	G A V L I M F W C
Pol	polar	Asp Glu Lys Arg His Ser Thr Asn Gln	D E K R H S T N Q
Ch	charged	Asp Glu Arg Lys His	D E R K H
Mp	MPCLVWIF	Met Pro Cys Leu Val Trp Ile Phe	M P C L V W I F
Ag	AGNQSTHY	Ala Gly Asn Gln Ser Thr His Tyr	A G N Q S T H Y
De	DEKR	Asp Glu Lys Arg	D E K R

Table 4.1 Legend of the abbreviations used on the x-axes in the figures 4.8 and 4.9

As a general observation regarding these results, is that none of the so-called predictive thermostability measures have strong predictive capabilities and the most obvious explanation is that amino acid composition itself is not the major factor contributing protein stability.

4.6. Relationship between thermostability and protein class

The Structural Classification of Proteins database (SCOP) comprehensively organizes all proteins with known structures based on their evolutionary and structural relationships (23). In order to address the possible correlation between protein thermostability and structure, we compared the structures from the SCOP database in respect to our thermostable protein grouping. SCOP classification is organized on

hierarchical levels: class, fold, superfamily, and family. Within those, superfamilies and families are said to have a common fold if their proteins have the same major secondary structures in the same arrangement with the same topological connections. Most of the folds are assigned to structural classes: all-α (SCOP class a), all-β (class b), α/β (class c), and $\alpha+\beta$ (class d).

Within our set of results, the classes are defined at the domain level, and if the protein has more than one domain, it belongs to the class defined by all of the present domains. Our results show that all-β protein class are significantly more frequent within the isolated and identified Thermally enhanced *S. solfataricus* proteins than the average genomic dataset. All-α proteins are also more frequent in this group. On the other hand Thermally enhanced *E. coli* proteins are mostly all-α and all-β with lesser frequency than the whole organism's proteome.

Rational for the found increase in highly stable beta proteins could be that a β barrel is found in many transcription proteins, including initiation and elongation factors, and also some ribosomal proteins, favoring its importance in preservation of essential cellular functions and therefore elevated stability properties, even though in these cases the fold is elaborated with additional structures. For example, beta barrel domain is represented within the elongation factors EF-Tu (24) and eEF1A (25) both of which function to recognize and transport aminoacyl-tRNA to the acceptor site of the ribosome during the elongation process, and of EF-G (26) which functions in translocating the peptidyl tRNA from the acceptor site to the peptidyl site. It is also present in initiation factors, in domain 2 of eIF2 gamma subunit (27) and domains 2 and 4 of IF2/eIF5B (28) both of which function to transport the initiator methionyl-tRNA to the ribosome. This beta barrel domain may be involved in interactions with the switch 2 region to stabilize the relative orientations of the domains, which undergo functionally important conformational changes between GTP- and GDP-bound states.

All-α class in our datasets has been statistically more frequent in Thermally selected *E. coli* protein subset and also in *S. solftaricus*, which

is in agreement with data from literature (17) that thermophilic proteins have a higher occurrence of residues in helical conformation. This result is not unusual; in fact this group contains some of the most abundant, highly stable proteins like ferredoxin. Two Fe4-S4 clusters present in the α-helical ferredoxin domain in *E. coli*. Iron-sulphur proteins have a significant function in electron transfer processes and in various enzymatic reactions. The α-helical ferredoxin domain is present in several proteins involved in redox reactions, including the C-terminal of the respiratory proteins succinate dehydrogenase (SQR) in bacteria/mitochondria, and fumarate reductase (QFR) in bacteria (29). Within the class of all-α are also ribosomal proteins of *E. coli* like Ribosomal protein S7 that is one of the proteins from the small ribosomal subunit.

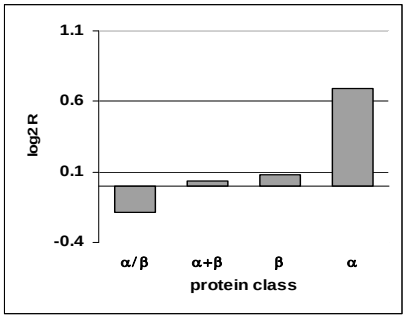


Fig. 4.10 Thermally enhanced sequences subset of *E. coli*: Frequency of protein class in respect to the complete proteome of the non treated organism.

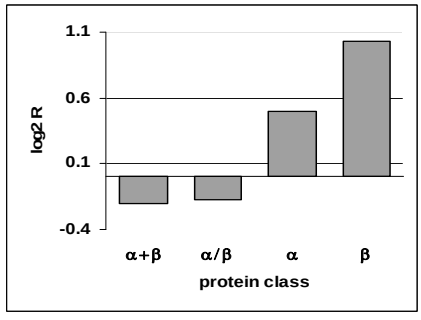


Fig. 4.11 Thermally enhanced sequences subset of *S. solfataricus*: Frequency of protein class in respect to the complete proteome of the non treated organism.

Protein class α+β is statistically more frequent only with *E. coli* Thermally enhanced group. This class of proteins is presented by mainly antiparallel β sheets with segregated α and β regions. Possible explanation may exist within the fact that in the *E. coli* cytosol, a fraction of the newly synthesized proteins requires the activity of molecular chaperones for folding to the native state that belong to this SCOP class. The major chaperones implicated in this folding process are the ribosome-associated Trigger Factor (TF), and the DnaK and GroEL chaperones with

their respective co-chaperones. Trigger Factor is an ATP-independent chaperone and displays chaperone and peptidyl-prolyl-cis-trans-isomerase (PPIase) activities *in vitro*, having the ability to interact with nascent chains as short as 57 residues renders TF as the first chaperone that binds to the nascent polypeptide chains (30). These groups of sequences contain the ribosomal subunit association domain.

In our results it is surprising that α/β protein class is much less frequent in Thermally enhanced subset than in the whole *E. coli* as well as *S. solfataricus* proteome, which is very unexpected and hard to explain, especially keeping in mind that important proteins in metabolic processes like histidine biosynthesis enzymes, tryptophan biosynthesis enzymes, etc.

4.7. Relationship between protein thermostability and cellular biological function - cellular thermo tolerance

It is difficult to say which biological cellular processes is more essential than the other when the life of a cell is based on the mutual functioning, overlapping and complementation of most of those processes. Therefore, we can only discuss which cellular processes have more of a thermostable character based on the functions of our identified protein sequences that were able to survive thermal treatment. Cellular functions being discussed here are not attributed to the identified sequences according to a particular fold but orthologous evolutionary connections.

The COGs database relies on phylogenetic classification of the all the proteins encoded in complete sequenced genomes of bacteria, archaea and eukaryotes, available at the web (http://www.ncbi.nlm.nih.gov/COG). The COG were constructed by applying the criterion of consistency of best hits specific to particular genome to the results of an exhaustive comparison of all protein sequences from those genomes. Due to the existence of orthologous relationships, orthologs are defined as the delineation of clusters of orthologous groups (COGs). Each COG consists of individual orthologous

genes or orthologous groups of paralogs from three or more phylogenetic lineages. Any two proteins from different lineages that belong to the same COG are orthologs and each COG is assumed to have evolved from an individual ancestral gene through a series of speciation and duplication events (31). Individual COGs are assigned to general functional categories, which represent major cellular processes, and in some cases, if known, to more specific pathways or systems. The COG functional categories, identified by one-letter codes, (table 4.2) are functional classification of genes conserved across different organisms with functions that are maintained and modified across phylogenetic groups during evolution.

Class	COG functional classification	Class	COG functional classification
A	RNA processing and modification	N	Cell motility
B	Chromatin structure and dynamics	O	Posttranslational modification, protein turnover, chaperones
C	Energy production and conversion	P	Inorganic ion transport and metabolism
D	Cell cycle control, mitosis and meiosis	Q	Secondary metabolites biosynthesis, transport and catabolism
E	Amino acid transport and metabolism	R	General function prediction only
F	Nucleotide transport and metabolism	S	Function unknown
G	Carbohydrate transport and metabolism	T	Signal transduction mechanisms
H	Coenzyme transport and metabolism	U	Intracellular trafficking and secretion
I	Lipid transport and metabolism	V	Defense mechanisms
J	Translation	W	Extracellular structures
K	Transcription	Z	Cytoskeleton
L	Replication, recombination and repair	-	Not in COGs
M	Cell wall/membrane biogenesis		

Table 4.2. COG functional classification categories

Our datasets representing the frequencies of the COG groups of the thermally selected soluble cytoplasmatic proteins from *S. solfataricus* and *E. coli* were analyzed and compared to the corresponding frequencies in the respect to the untreated organism proteome. This way we kept in mind that some proteins are present in the cell in the elevated amount compared to the others as the normal cellular property. Categories that were obviously enriched only due to the resistance to the applied thermal separation are explained from the point of view of the biological process in which they participate in the living system, within the ecological niche where organism naturally belongs.

In thermally treated cytosolic extract from both investigated organisms,

thermally enriched functional categories were: P – inorganic ion transport and metabolism, E - aminoacid transport and metabolism, K – transcription, T – signal transduction mechanism and O - chaperones, protein turnover and posttranslational modifications. Diminished frequency is noticed within the categories L - replication, recombination and repair, I – lipid transport and mechanism, M - wall membrane biogenesis and R – group with general function prediction only.

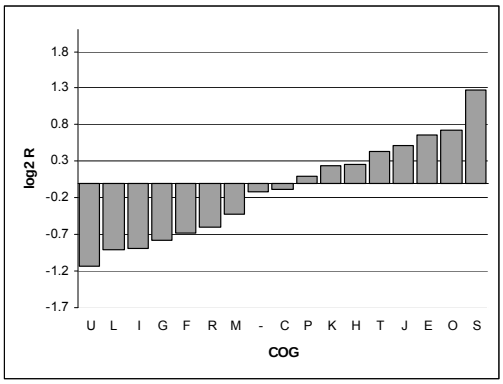


Fig.4.12 Thermally enhanced sequences subset of *E. coli*: Frequency of COGs in respect to the complete proteome of the non treated organism.

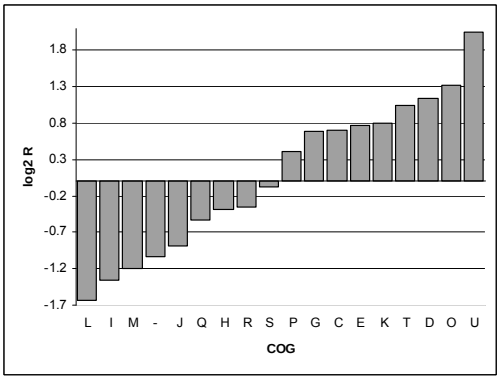


Fig. 4.13 Thermally enhanced sequences subset of *S. solfataricus*: Frequency of COGs in respect to the complete proteome of the non treated organism

Other functional categories were enriched or diminished either in the

bacterial or in the archaeal Thermally enhanced cellular proteome. Interestingly, functional category U – intracellular trafficking and secretion was the category with the most diminished frequency in mesophile *E. coli*, but the most increased in hyperthermophile *S. solfataricus*. According to this, would be rational to conclude that this group of processes needs to have highly thermostable participants in thermophilic organism, which is the only way to preserve proteins about to be secreted in to harsh extracellular environment. This property is less important with mesophile, as its environment does not impose serious themperature component.

Speaking of essential cellular processes, life preserving mechanism may give the explanation why these categories were enhanced either in stability or at least in quantity. The key player in the translation process, the ribosome, is accompanied by RNA polymerases and RNA-handling proteins. Number of other processes may additionally be needed for a successful translation. These back-up processes may provide the necessary components for the synthesis of ribosomal/messenger RNA within sugar and nucleotide metabolism, and polypeptide chains for amino-acid metabolism. Both cytoplasmic and mitochondrial translation machineries still require the same sugar and nucleotides as building bricks for transcription and amino acids for translation. This metabolic network is being coupled to the cytoplasmic machinery. As we are dealing only with cytosolic soluble proteome in this study, only the protein functional categories from this cellular compartment may have been, and indeed were identified.

Another explanation for overlapping some of these functional categories from the two such distinct organisms may be given taking advantage of the operon organization of their genomes. An operon is made up of genes that are transcribed as part of a single mRNA molecule. In both bacterial and archaeal organisms co-transcribed genes are co-regulated at the transcriptional level and often have related roles, for example involving protein-protein interactions or as part of the same metabolic pathway. COG functional categories are universal and very similar in different organisms suggesting also that protein with the same function, but not necessarily evolutionarily related, could replace each other in

different organisms. In bacteria, in addition to some Hsp70-like chaperones, the trigger factor is associated with the ribosome (32) being involved in co-translational protein folding. Extending this to other type of organism like archaeal suggests that there may be a general mechanism to facilitate protein folding. Protein translation and folding are housekeeping processes and its components are constitutively expressed.

Global occurrence of archaeal species nowadays is very diverse as well as their living environment, but ancient archaeal organisms were exposed to hot habitat. During evolution in such kind of environment, archaeal proteins were initially designed in a hot environment in a way that its inherent structural properties enable thermal resistance with sequences able to fold and be stable in such thermostable structures. Alternatively, bacterial species that evolved later, initially as a mesophilic organism that only later recolonized a hot environment were involved in secondary thermophilic adaptation that required the enhancement of the thermostability of already existing proteins. Comparative analysis of structures and complete genomes of several hyperthermophilic archaeal organisms (e.g. *Pyrococcus furiosus*) and bacterial (33) (e.g. *Thermatoga maritima*) revealed that organisms develop diverse strategies of thermophilic adaptation by using these two fundamental physical mechanisms of thermostability.

Orthologues are genes that evolved from a single ancestral gene in the last common ancestor of the compared genomes, while paralogues are genes that are related by duplication (34). Convergent evolution represents a phenomenon when two distinct species with differing ancestries evolve to display similar physical feature (35). Environmental circumstances that require similar developmental or structural alterations for the purposes of adaptation can lead to convergent evolution even though the species have different origin. As a consequence of convergent evolution, biological structures or species that exhibit similar functions or/and appearance may appear, even though they evolved through widely divergent evolutionary pathways and had different ancestors. These similarities are typically explained as the result of common adaptive solutions to similar environmental

pressures on the level of the organism.

We may speculate that these processes were in base of choosing our identified thermostable protein datasets and respective biological processes as the ones that evolutionary needed more secure properties in order to enable the organism to perform its activities in the thermally challenging situations.

4.8. Conclusions

Exhaustive thermal treatment of soluble cytosolic proteome has been a good metod of isolating highly stable proteome subset. Soluble cytosolic proteome of *S. solfataricus* as a representative of the extremophiles and a known source of wide range of thermostable proteins for scientific studies has been compared to the one originating from *E. coli*, a model mesophilic organism. It was found that *per se* presence or absence of certain type of amino acid residues or group is not a sufficient prerequisite for difference in stability, bur rather the position and subsequent interaction of the residues with the surrounding ones can lead towards alteration in stability properties which is in agreement with various other studies. SCOP folds are also not very strongly biased in respect to the thermostability. As a general observation regarding this results, is that none of the so-called predictive thermostability measures have strong predictive capabilities and the most obvious explanation is that their combination may or may not lead towards enhanced thermal stability, which is influenced by the organisms' ecological surroundings and natural habitat that evolutionary select appropriate stability determinants. Rising scientific interest in the thermostability basics of proteins gave origin to many studies so far, and ranges of factors contributing to protein thermostability have been identified. However, no general mechanism that can be named as the most important factor for increased thermostability was found.

4.9. References

1. **She, Q., R. K. Singh, F. Confalonieri, Y. Zivanovic, G. Allard, M. J. Awayez, C. C. Chan-Weiher, I. G. Clausen, B. A. Curtis, A. De Moors, G. Erauso, C. Fletcher, P. M. Gordon, I. Heikamp-de Jong, A. C. Jeffries, C. J. Kozera, N. Medina, X. Peng, H. P. Thi-Ngoc, P. Redder, M. E. Schenk, C. Theriault, N. Tolstrup, R. L. Charlebois, W. F. Doolittle, M. Duguet, T. Gaasterland, R. A. Garrett, M. A. Ragan, C. W. Sensen, and J. Van der Oost.** 2001. The complete genome of the crenarchaeon Sulfolobus solfataricus P2. Proc Natl Acad Sci U S A **98:**7835-40.
2. **Chong, P. K., and P. C. Wright.** 2005. Identification and characterization of the Sulfolobus solfataricus P2 proteome. J Proteome Res **4:**1789-98.
3. **Blattner, F. R., G. Plunkett, 3rd, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of Escherichia coli K-12. Science **277:**1453-74.
4. **Durfee, T., R. Nelson, S. Baldwin, G. Plunkett, 3rd, V. Burland, B. Mau, J. F. Petrosino, X. Qin, D. M. Muzny, M. Ayele, R. A. Gibbs, B. Csorgo, G. Posfai, G. M. Weinstock, and F. R. Blattner.** 2008. The complete genome sequence of Escherichia coli DH10B: insights into the biology of a laboratory workhorse. J Bacteriol **190:**2597-606.
5. **Bergthorsson, U., and H. Ochman.** 1998. Distribution of chromosome length variation in natural isolates of Escherichia coli. Mol Biol Evol **15:**6-16.
6. **Hayashi, K., N. Morooka, Y. Yamamoto, K. Fujita, K. Isono, S. Choi, E. Ohtsubo, T. Baba, B. L. Wanner, H. Mori, and T. Horiuchi.** 2006. Highly accurate genome sequences of Escherichia coli K-12 strains MG1655 and W3110. Mol Syst Biol **2:**2006 0007.
7. **Ross, P. L., Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson, and D. J. Pappin.** 2004. Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics **3:**1154-69.
8. **Choe, L., M. D'Ascenzo, N. R. Relkin, D. Pappin, P. Ross, B. Williamson, S. Guertin, P. Pribil, and K. H. Lee.** 2007. 8-plex quantitation of changes in cerebrospinal fluid protein expression

in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease. Proteomics **7:**3651-60.

9. **D'Ascenzo, M., L. Choe, and K. H. Lee.** 2008. iTRAQPak: an R based analysis and visualization package for 8-plex isobaric protein expression data. Brief Funct Genomic Proteomic **7:**127-35.
10. **Einhauer, A., and A. Jungbauer.** 2002. Recombinant autofluorescent landmarks for standardization of electrophoretic migration of proteins. Electrophoresis **23:**1146-52.
11. **Burgess-Cassler, A., J. J. Johansen, D. A. Santek, J. R. Ide, and N. C. Kendrick.** 1989. Computerized quantitative analysis of coomassie-blue-stained serum proteins separated by two-dimensional electrophoresis. Clin Chem **35:**2297-304.
12. **Nandi, S., N. Mehra, A. M. Lynn, and A. Bhattacharya.** 2005. Comparison of theoretical proteomes: identification of COGs with conserved and variable pI within the multimodal pI distribution. BMC Genomics **6:**116.
13. **Schwartz, R., C. S. Ting, and J. King.** 2001. Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. Genome Res **11:**703-9.
14. **Kannan, N., and S. Vishveshwara.** 2000. Aromatic clusters: a determinant of thermal stability of thermophilic proteins. Protein Eng **13:**753-61.
15. **Kelkar, D. A., and A. Chattopadhyay.** 2006. Membrane interfacial localization of aromatic amino acids and membrane protein function. J Biosci **31:**297-302.
16. **Warren, G. L., and G. A. Petsko.** 1995. Composition analysis of alpha-helices in thermophilic organisms. Protein Eng **8:**905-13.
17. **Kumar, S., C. J. Tsai, and R. Nussinov.** 2000. Factors enhancing protein thermostability. Protein Eng **13:**179-91.
18. **Cambillau, C., and J. M. Claverie.** 2000. Structural and genomic correlates of hyperthermostability. J Biol Chem **275:**32383-6.
19. **Pace, C. N., B. A. Shirley, M. McNutt, and K. Gajiwala.** 1996. Forces contributing to the conformational stability of proteins. Faseb J **10:**75-83.
20. **Pack, S. P., and Y. J. Yoo.** 2004. Protein thermostability: structure-based difference of amino acid between thermophilic and mesophilic proteins. J Biotechnol **111:**269-77.
21. **Farias, S. T., and M. C. Bonato.** 2003. Preferred amino acids and thermostability. Genet Mol Res **2:**383-93.

22. **Zeldovich, K. B., I. N. Berezovsky, and E. I. Shakhnovich.** 2007. Protein and DNA sequence determinants of thermophilic adaptation. PLoS Comput Biol **3:**e5.
23. **Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia.** 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol **247:**536-40.
24. **Andersen, G. R., S. Thirup, L. L. Spremulli, and J. Nyborg.** 2000. High resolution crystal structure of bovine mitochondrial EF-Tu in complex with GDP. J Mol Biol **297:**421-36.
25. **Andersen, G. R., L. Pedersen, L. Valente, I. Chatterjee, T. G. Kinzy, M. Kjeldgaard, and J. Nyborg.** 2000. Structural basis for nucleotide exchange and competition with tRNA in the yeast elongation factor complex eEF1A:eEF1Balpha. Mol Cell **6:**1261-6.
26. **Laurberg, M., O. Kristensen, K. Martemyanov, A. T. Gudkov, I. Nagaev, D. Hughes, and A. Liljas.** 2000. Structure of a mutant EF-G reveals domain III and possibly the fusidic acid binding site. J Mol Biol **303:**593-603.
27. **Schmitt, E., S. Blanquet, and Y. Mechulam.** 2002. The large subunit of initiation factor aIF2 is a close structural homologue of elongation factors. Embo J **21:**1821-32.
28. **Roll-Mecak, A., C. Cao, T. E. Dever, and S. K. Burley.** 2000. X-Ray structures of the universal translation initiation factor IF2/eIF5B: conformational changes on GDP and GTP binding. Cell **103:**781-92.
29. **Iverson, T. M., C. Luna-Chavez, L. R. Croal, G. Cecchini, and D. C. Rees.** 2002. Crystallographic studies of the Escherichia coli quinol-fumarate reductase with inhibitors bound to the quinol-binding site. J Biol Chem **277:**16124-30.
30. **Deuerling, E., H. Patzelt, S. Vorderwulbecke, T. Rauch, G. Kramer, E. Schaffitzel, A. Mogk, A. Schulze-Specking, H. Langen, and B. Bukau.** 2003. Trigger Factor and DnaK possess overlapping substrate pools and binding specificities. Mol Microbiol **47:**1317-28.
31. **Tatusov, R. L., E. V. Koonin, and D. J. Lipman.** 1997. A genomic perspective on protein families. Science **278:**631-7.
32. **Fink, A. L.** 1999. Chaperone-mediated protein folding. Physiol Rev **79:**425-49.
33. **Nelson, K. E., R. A. Clayton, S. R. Gill, M. L. Gwinn, R. J. Dodson, D. H. Haft, E. K. Hickey, J. D. Peterson, W. C. Nelson, K. A. Ketchum, L. McDonald, T. R. Utterback, J. A. Malek, K. D. Linher, M. M. Garrett, A. M. Stewart, M. D. Cotton, M. S. Pratt, C. A.**

Phillips, D. Richardson, J. Heidelberg, G. G. Sutton, R. D. Fleischmann, J. A. Eisen, O. White, S. L. Salzberg, H. O. Smith, J. C. Venter, and C. M. Fraser. 1999. Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of Thermotoga maritima. Nature **399:**323-9.

34. **Gogarten, J. P., and L. Olendzenski.** 1999. Orthologs, paralogs and genome comparisons. Curr Opin Genet Dev **9:**630-6.
35. **Stevenson, J. C.** 1991. Dictionary of concepts in physical antropology. Greenwood Press, New York.

The results presented in this chapter resulted of collaboration with following laboratories and people:

Collaboration with Environmental Systems Group, Department of Chemical and Process Engineering, University of Sheffield, UK, headed by Prof Phillip Wright. iTRAQ experiments were performed in collaboration with Dr Trong Khoa Pham during a visit period in November-December 2008.

Collaboration with Computational Genomics Laboratory, Instituto Gulbenkian de Ciência, Oeiras, Portugal headed by Dr José B. Perreira Leal. Member of this group, Renato Alves, provided computational algorithm for data processing.

Chapter 5

RELATIONSHIP BETWEEN PROTEIN THERMOSTABILITY AND SOLUBILITY IN *Escherichia coli* THERMALLY SELECTED SUBPROTEOME

Relationship between protein thermostability and solubility in *Escherichia coli* thermally selected subproteome

5.1. Summary

In the presented work, we correlate the determinants underlying the thermostability and aggregation properties of the thermally isolated subset of proteins that are present in the bacterial cytosol. Thermally isolated (90^{o}C) cytosolic proteome subset from *Escherichia coli* was identified and further investigated for different physicochemical characteristics and parameters such as isoelectric point, molecular weight, aliphatic index, grand average hydropathy (GRAVY) and also correlation with protein structure. Importance of the identified sequences in the life-preserving processes in the cell was also taken into account. Computed results were compared to available data from the literature regarding correlation of these properties from our "super-thermostable" subproteome and other, namely highly soluble one, leading to an interesting conclusion that the identified determinants are in fact rather common. A development of an approach to mutually screen for thermostability and solubility properties in cellular environment may insert improvement into progression of biotechnological processes that highly depend on protein stability and/or modulation of misfolding and aggregation, with a hope to upgrade processes addressed.

5.2. Introduction

This work was focused onto identifying common basis of protein stability to high temperatures and solubility properties, with an overview of possible relationship of protein structure to function, and the evolutionary relationships between them.

Intracellular ambient and chaperone function

Intracellular environment of living cells is highly crowded due to high concentration of small molecules, macromolecules and supramolecular assemblies, between ~50-400 mg/ml. The macromolecules present in the cell occupy about 40% of the total medium volume, therefore accessible

volume in the cell is reduced, significant fraction of the water is involved in solvation and does not behave as bulk water. Molecular crowding has a complex effect on interactions between all types of biological molecules and the rate of biochemical reactions, where the overall result of the present factors depends on the specific nature of each reaction (1). Protein folding in the living cell is a process determined by amino acid sequence (2), influenced by specific biological circumstances in the living cell, and often assisted by the presence of chaperone system for the correct folding of nascent polypeptide chains (3, 4). Some proteins are able to assume their native state unassisted, but for some the presence of molecular chaperones is essential in supporting the assembly of protein structures in a biologically relevant timescale, without being present in the final functional form of these structures (5). During folding process or translocation, chaperones participate in shielding hydrophobic surfaces of an incomplete protein preventing the binding of the polypeptide chains with each other and subsequent aggregation (3). Additionally, chemical attractions and interactions between molecules also influence the final protein product. For soluble, globular proteins in particular, the energy required to bury hydrophobic residues is one of the main driving forces behind the folding of the polypeptide molecule and subsequent stability (6). Once folded, hydrophobic interactions are essential in retaining the tertiary structure of the protein. Such hydrophobic regions may indicate a buried core within the protein structure, or a feature like e.g. a transmembrane segment.

Solubility of globular proteins in intact cellular environment is preserved by coordinated involvement of a vast range of quality control mechanisms. Various scientific studies have been devoted to investigating and predicting the propensity of a protein to be stable and soluble on overexpression, especially regarding heterologous expression in *E. coli*, and developing methodologies in order to predict solubility/aggregation propensities based on aminoacid sequence (7-10). Obtaining sufficient soluble proteins is a frequent limiting factor in experimental studies as well as technological processes. Stability and solubility alterations and consequent aggregation of certain proteins are of a great interest in investigating the diseases that are triggered by

protein aggregation (11). Investigating stability properties of highly stable proteins, especially thermostable ones have an emerging interest both from fundamental point of view, in studies aiming at the elucidation of the stability determinants (12, 13) or protein structure e.g. (14-16), and also from an applicative point of view, in biotechnological or industrial processes where a biological catalyst is used (15, 17).

This work was focused onto identifying the basis of protein elevated thermostability as integral to our understandings of protein folding, the relationship with function, and the evolution, that were discussed in previous chapters. Investigating determinants of thermostability pointed to additional aspect of protein solubility that was intriguing enough to attempt an additional perspective on thermostable cellular proteins.

Thermostable soluble proteome subset

Our previous studies regarding highly stable cytosolic soluble proteins (16) originating from thermostable archaeal organism from the order *Sulfolobales* have gave an insight into mechanisms and determinants of protein thermostability, but also some additional topics for discussion from ecological and evolutionary points of view. We have discussed that essential life preserving processes in the living cell are likelier to involve more stable protein sequences in order to protect the life itself and that exhaustive thermal treatment of soluble cytosolic proteome has been a good method of isolating highly stable proteome subset.

These results were in agreement with data from literature stating that *per se* presence or absence of certain type of amino acid residues or group is not a sufficient prerequisite for difference in stability, bur rather the position and subsequent interaction of the residues with the surrounding ones can lead towards alteration in stability properties. SCOP folds are also not very strongly biased in respect to the thermostability. General observation regarding these results is that none of the so-called predictive thermostability measures in fact have strong predictive capabilities and the most obvious explanation is that their combination may or may not lead towards enhanced thermal stability. As a continuance of this investigation, further study provided more results and developed conclusions based on isolation and identification of highly

thermostable proteins originating from mesophilic bacterial organism *E. coli* as the model system, using the same initial methodology, exhaustive thermal treatment of soluble cytosolic proteome in order to isolate highly stable proteome subset.

Sequences that were identified as highly stable after thermal treatment (90^{o}C) of *E. coli* cytosolic proteome, were cross-compared with the results published by the group of Taguchi and co-workers regarding bimodal solubility/aggregation distribution of *E. coli* proteome (18) of the proteins expressed in cell-free, chaperone-free system named PURE. The intersection of those two groups of sequences - superstable cytosolic proteins that survived harsh thermal treatment remaining soluble, and chaperone-free-expressed soluble proteins, was investigated in order to check for common determinants between these two important properties: thermostability and solubility. Interesting conclusions emerged, guiding towards the opinion that cellular guidelines that favor protein thermostability are in common ones that favor solubility, or lack of propensity to aggregate, of important life preserving sequences.

5.3. Objectives and Methodologies

There are several published studies regarding correlation of protein solubility and other sequence-derived properties. Wilkinson and Harrison (19) proposed a method that was latter improved for calculating protein solubility from the known sequence, that is based on parameters: average charge, determined by the relative numbers of Asp, Glu, Lys and Arg residues, and the content of turn-forming residues (Asn, Gly, Pro and Ser). It was latter demonstrated that insoluble proteins tended to have more hydrophobic stretches (longer then 20 amino acids), lower glutamine content (Q<4%), fewer negatively charged residues (DE<17%) and higher percentage of aromatic amino acids (FYW>7.5%) than soluble ones (20), allowing prediction of protein solubility with 65% accuracy. Also, high content of negative residues (DE>18%) and absence of hydrophobic patches are associated with improved solubility and also low percentage of aspartic acid, glutamic acid, asparagines and glutamine residues (DENQ<16%) increases the probability of a protein to be insoluble. Analysis of more than 27 000 proteins from multiple organisms

found that protein solubility is influenced by (in decreasing order of importance): percentage of serine (S <6.4%), fraction of negatively charged residues (DE <10.8%), percentage of S, C, T and M amino acids, and length (<516 amino acids), in decreasing order of importance (21).

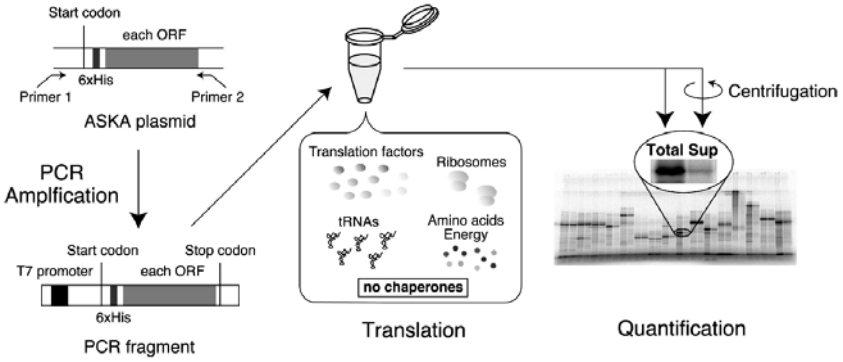


Fig. 5.1. Schematic illustration of the PURE expression system (18). Each ORF in the ASKA library, which has all of the *E. coli* ORFs, was amplified by PCR using two common primers to translate the gene in the cell-free translation system. The reconstituted cell-free translation system (the PURE system) contains no chaperones. After the 60-min translation, an aliquot of the translation mixture was centrifuged to obtain the soluble fraction. The uncentrifuged (Total) and supernatant (Sup) fractions were subjected to SDS/PAGE, and the translated products were quantified by autoradiography (18).

In the continuance with the results presented in the previous chapter of this thesis, this study further developed results and conclusions based on isolation and identification of highly thermostable proteins originating from mesophilic bacterial organism *E. coli* as the model system. Process of analysis started with exhaustive thermal treatment of soluble cytosolic proteome in order to isolate highly stable proteome subset. Soluble cytosolic extract was exposed to high temperature up to 90^{o} C during prolonged period of time, less stable denatured proteins were removed by bench centrifugation, while proteins that remained soluble after this treatment were subjected to further analysis. iTRAQ protein sequences identification and quantification of this "thermoproteome" and bioinformatics analysis gave an insight into possible relationship between

thermostability on one side and various other important protein properties. Our results were subsequently compared to the results recently published in literature (18, 22, 23), revealing accordance with the most of the information present in the literature up to date in following manner: Additional point of discussion was comparing our data regarding the *E. coli* cytosolic proteome subset with elevated thermostability, with the information present in the literature regarding the *E. coli* cytosolic proteome subsets with highly pronounced solubility properties.

Taguchi and co-workers (18) have previously published interesting set of results that can relate with our findings in a mutually supporting manner. They used a model named PURE, a reconstituted system (24) containing only the essential *E. coli* factors responsible for protein synthesis (Fig 5.1, (18)). Complete ASCA library - consisting of all ORFs - was translated, but in the presence only of the essential *E. coli* factors responsible for protein synthesis and excluding the presence of the chaperones. They performed a comprehensive analysis, successfully quantified more than 3000 protein sequences of the initial 4000 ORFs. The fact that this cell-free translation system contains no chaperones enabled to investigate and evaluate inherent aggregation propensities (14, 17). They examined propensity for protein aggregation by a centrifugation assay where an aliquot of the translation mixture was centrifuged and the solubility was determined as the proportion of the supernatant fraction obtained after the centrifugation of the translation mixture, to the uncentrifuged total protein. They have also invested a considerable effort in identifying residues stabilizing the native conformations of proteins. The aggregation propensities of proteins, which were evaluated under the chaperone-free condition, showed that the proteins were categorized into two groups, soluble and aggregation-prone. Another conclusion was that some of the SCOP folds are strongly biased to the aggregation propensity which is apparently paradoxical because aggregates formation should occur before the completion of folding. We have cross-compared our results of the thermally (90^{o}C) selected *E. coli* cytosolic proteome subset with the results published by Taguchi's group regarding bimodal solubility/aggregation distribution of *E. coli* proteome(18) publicly

available at http://www.taguchi.bio.titech.ac.jp/eng/paper-e/assets/2009_PNAS_ecoli_proteins_solubility.xls .

Our subset of "superstable" cytosolic proteins that survived harsh thermal treatment and still remained in the solution, when compared to the published ones, obtained results available in this segment of this thesis.

Publically available PubMed, at the National Center for Biotechnology Information (NCBI) http://www.ncbi.nlm.nih.gov was used to access the bibliographic database. To check for easily detectable biases in our data we evaluated the classification performance of simple global sequence dependent protein features (isoelectric point, molecular weight, aliphatic index and GRAVY index) by retrieving them from the ProtParam tool of the ExPASy proteomics server of the Swiss Institute of Bioinformatics http://us.expasy.org/tools/protparam.html. Information regarding COG functional categories was retrieved from http://www.ncbi.nlm.nih.gov/COG/old/xognitor.html

5.4. Results and discussion

In all of the presented results in this chapter, abbreviation "*S*" refers to the group of sequences that are result of overlapping data from our thermally enhanced *E. coli* sequences with results from Taguchi and co-workers (18). The intersection of those two groups of sequences: a) superstable cytosolic proteins that survived harsh thermal treatment remaining soluble, and b) chaperone-free-expressed soluble proteins, was investigated in order to check for common determinants between these two important properties: thermostability and solubility. Sequences "S" data are available in Appendix III.

Solubility predictions

A statistical model for prediction of solubility on expression in *E. coli* defined by Wilkinson and Harrison shown to be useful in the selection of proteins with high solubility (19) or to estimate the solubility of proteins expressed in *E. coli*. They used discriminant analysis to compare proteins according to six composition related parameters: charge average, turn forming residue fraction, cysteine fraction, proline fraction, hydrophilicity

and total number of amino acid residues. The relative number of turn forming residues (asparagine, glycine, proline and serine) and absolute charge per residue (fraction of positively and negatively charged amino acids) correlate strongly with inclusion body formation. A composite parameter (CV-canonical variable) is dependent on the contribution of each of the individual amino acid:

$$CV = 15.43\{(N+G+P+S)/n\} - 29.56\{[(R+K)-(D+E)/n] - 0.03\}$$

where N, G, P, S, R, K, D, E are the absolute numbers of asparagine, glycine, proline, serine, arginine, lysine, aspartic acid and glutamic acid residues, respectively, and n is the total number of residues in the whole sequence. A threshold discriminate CV' = 1.71 distinguish soluble proteins from insoluble ones. A protein is predicted to be soluble, if the difference between CV and CV' is negative. On the contrary, a CV-CV' difference larger than zero, predicts the protein to be insoluble. A probability of solubility is calculated (25) :

$$P = 0.4934 + 0.276(CV - CV') - 0.0392(CV - CV')^2$$

The higher the absolute of CV-CV', the higher the probability of solubility (CV-CV' < 0) or insolubility (CV-CV' > 0). CV-CV' values of -0.4, 0.0 and 1.1 indicate probabilities of solubility of 60 %, 50 % and 25 %, respectively. Calculating probability of solubility according to this model is facilitated by publically available web site http://www.biotech.ou.edu/ where only input requirement is amino acid sequence.

In our dataset "*S*", overall probability of solubility is remarkable 83.5%. Keeping in mind that we are dealing with a thermally isolated cytosolic proteome subset, with an outstandingly high overall solubility, we may emphasize strong correlation between thermal stability and solubility. Relation between intrinsic protein thermostability and life preserving cellular processes has been discussed previously (16). In that context, proteins executing important cellular functions tend to be better adapted and less prone to against aggregation than nonessential ones, suggesting an intrinsic evolutionary mechanism to preserve normal cellular physiological functions. Consequences of protein aggregation on cellular

function would be ultimately associated to either decrease in efficiency or lack of function of certain cellular metabolic processes that would consequently lead to alteration or/and end of physiological life preserving pathways. Therefore, it is conceivable that evolution selects an overall decreased aggregation propensity in evolutionary essential proteome subsets.

Bimodal solubility distribution of thermostable proteins

Protein solubility is a strong evolutionary constrain, in addition to protein function. Any protein can remain functional in its native state under physiological conditions at its specific cellular localization. Results indicate that Aggregation propensity is not evenly distributed across the overall group of our thermally isolated cytosolic proteome subset (Fig. 5.2) with obvious bimodal distribution. Proteins that are aggregation prone (Agg) are defined as sequences with solubility properties up to 30%, and highly soluble (Sol) defined as over 70% solubility protein groups.

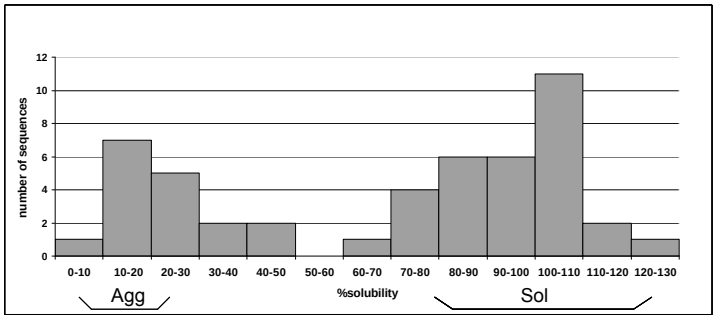


Fig 5.2 Bimodal solubility distribution of sequences within "S" dataset

Same results have been found in the work of Taguchi and co-workers (18), where the data were based on results on much larger dataset, more than three thousand *E. coli* translated protein sequences. Keeping in mind that this group was dealing with thermally untreated completed proteome, and that the results were the same even after subtracting the data referring to membrane proteins, it is obvious that this type of solubility property distribution is universal both for treated as well as untreated cellular proteome. It is generally accepted that higher protein

concentrations generate more protein aggregates (26), but this is not the case, because there is no apparent correlation between the solubilities and the yields. In our dataset proteins were not individually quantified, but the concentration of the *E. coli* soluble subfraction of the cellular extract after the 90^{o}C treatment and subsequent centrifugation was far beyond the trigger limit of protein concentration that would further influence solubility.

What is interesting to underline here, is that our data set included overlap of our isolated the most thermally stable cytosolic proteins with Taguchi's overall dataset, and the same distribution was maintained, but our "S" proteome subgroup had medium probability of solubility of remarkable 83.5%. This emphasizes the previous conclusion of strong correlation between thermal stability and solubility, while maintaining bimodal solubility property distribution.

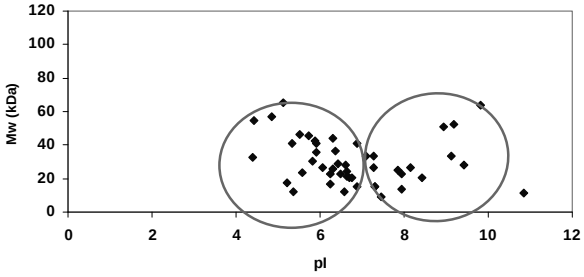


Fig. 5.3. Correlation between pI and MW, bimodal distribution.

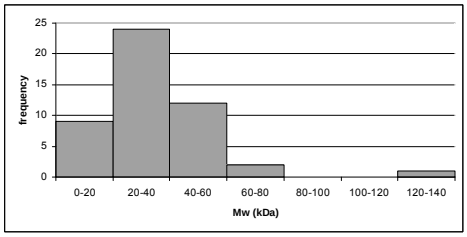
It is also important to note that Sol solubility subgroup is presented with 62% sequences of our initial highly thermostable "S" dataset, with mean solubility of amazing 96.4%, which further underlines common selection trend towards protection of the sequences with highly elevated thermal stability and solubility within the cell. The heat stability and solubility of such proteins can be explained by specific features inherent to their structures, or it can be acquired by evolution in order to satisfy the needs of performing specialized functions. Properties that can be responsible for such behavior are further investigated and represented in this chapter with an additional outline of functional categories that they belong to.

Solubility correlation with pI/Mw

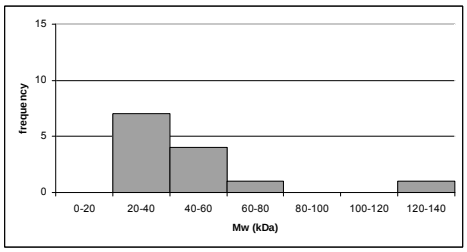
We compared the physicochemical properties of the proteins, such as the molecular mass and the deduced isoelectric points (pI), to address the relationship between solubility, themal stability and sequence. One should keep in mind that considered MW and pI are theoretical, not experimental, according to amino acid sequence of the specific protein. The molecular weight was added because it correlates better with size than the number of residues. The molecular weight of each protein was determined with aid of the pI/MW tool from the Swiss Institute of Bioinformatics http://web.expasy.org/compute_pi/.

Small proteins, below 50 kDa are represented with 85% of sequences in our dataset. In literature bimodal pI distribution exists in prokaryotic proteome (27) with peaks centered around pH 5.5 and pH 9, and in our case around pH 6,1 and pH 8,7 (Fig. 5.3). This bimodality was explained as being caused by the fact that being least soluble at their pI, proteins have evolved to have pI values away from neutral pH – which was assumed to be the intracellular pH. Our data are consistent with this, with obvious bimodal distribution around mentioned pH. Other studies discussing protein pI in correlation with MW and other global sequence parameters, or solubility properties also failed to prove direct connection (8, 19, 28, 29) whether the sequences are originally present in the natural quantities or overexpressed. It should be noted that our data set included isolated the most thermally stable cytosolic proteins with, and the same distribution was maintained. Further in the literature, the presence of a trimodal pI distribution is observed in correlation of pI to intracellular localization: cytoplasmic, nuclear and membrane proteins seemed to lie largely in the acidic, neutral and basic portions of the trimodal distribution, respectively (30), but in our case cellular compartmentalization is not of interest as we are dealing only with small fraction of soluble cellular proteome subset.

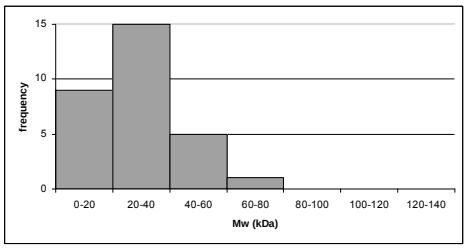
Proven bimodal distribution of our results is an interesting property, as it correlates thermostability or solubility of our isolated subproteome. On the other hand it is not an exclusive property. Literature data focused more on an evolutionary approach, molecular weight of proteins was found to be much more conserved feature than their pI value as a lot of orthologous proteins change their pI between acidic and basic and only a few stay exclusively acidic or basic in different organisms (31), and most of the proteins with thermostable properties are expected to be small, like in our dataset where the most of the sequences are in the range of 20-40 kDa. In fact, average MW was 31,8 kDa within complete "S" dataset, that is 44,2 kDa and 26,7 kDa within Agg and Sol subsets respectively (Fig. 5.4 a), b) and c)). Taking all of it into account, bimodal pI protein distribution and small molecular weight can not directly explain and define the relationship between solubility and thermal stability within our dataset, but rather accompany data already present in the



a) Molecular mass in Total "S"



b) Molecular mass in Agg subgroup



c) Molecular mass in Sol subgroup

Fig. 5.4. Correlation between solubility and physicochemical properties. Histograms of molecular mass in the a)Total, b) Agg, and c) Sol

literature.

Relationship between solubility, aliphatic index and GRAVY index of thermostable proteins

The aliphatic index (AI) of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins and is calculated according to the following formula (32)

$$AI = X(Ala) + a \times X(Val) + b \times [X(Ile) + X(Leu)]$$

where X(Ala), X(Val), X(Ile), and X(Leu) are mole percent (100 X mole fraction) of alanine, valine, isoleucine, and leucine. The coefficients a and b are the relative volume of valine side chain (a = 2.9) and of Leu/Ile side chains (b = 3.9) to the side chain of alanine.

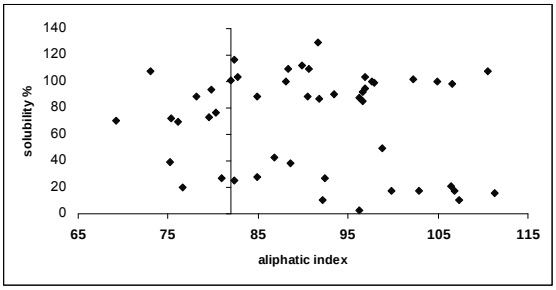


Fig 5.5 Correlation between solubility and aliphatic index in “S” dataset

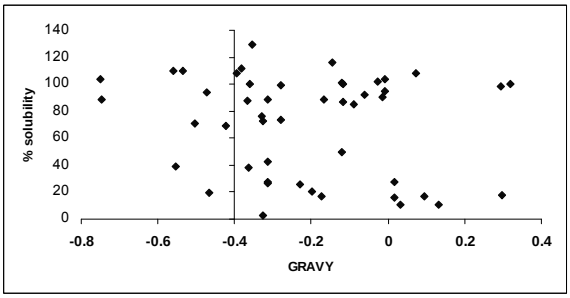


Fig. 5.6. Correlation between solubility and GRAVY index in “S” dataset

The AI of proteins from thermophilic bacteria was found to be significantly higher than that of ordinary proteins and can serve as a measure of thermostability of proteins (32). But, in our study, source of our protein sequences are not thermophilic bacteria that are evolutionary adapted to the lifestyle under elevated temperature conditions, but carefully selected proteome subset originating from mesophilic bacterium, proven to be highly thermostable, and now checked for solubility properties by overlapping our dataset and the publicly available information on highly soluble cytosolic sequences (18) from Taguchi and co-workers. In their PURE model the cell-free translation system that only contains the essential *Escherichia coli* factors responsible for protein synthesis but no chaperones, so that inherent aggregation propensities can be evaluated.

Thermolabile folding intermediates have been suggested to contribute to inclusion body formation by exhausting the in vivo supply of chaperonins, as they serve as chaperonin substrates (33). Also, mean value for aliphatic index found in literature for proteins find to be insoluble in *E. coli* was 82, while higher aliphatic index is present in thermostable, soluble sequences, which suggest that an increase in the thermostability of the protein might favor an increase in its solubility (29). In our "S" dataset 75% of sequences have aliphatic index higher than 82 (Fig 5.5). Higher aliphatic index for our dataset proteins, imply that an increased thermostability of the proteins may support an increase in their solubility and again emphasize the conclusion that thermostability and solubility have a positive correlation.

Peptides grand average hydropathicity (GRAVY) indicates the hydrophilicity or hydrophobicity of a protein, and can be calculated as an arithmetic mean of the sum of the hydropathy index of each amino acid of a protein divided by the number of residues in the sequence. The GRAVY score was calculated using the publicly Web-available application http://web.expasy.org/protparam/ (34). Hydropathy score values take into account both the number of hydrophobic and the number of hydrophilic residues and may be represented by a positive or a negative score. Hydropathy is a parameter that is used to determine the hydrophobicity of various regions of the molecule. Hydropathy scale is

widely used to obtain a measure of the effective interaction between any two amino acid residues in proteins and is based on the assumption that attraction between two hydrophobic groups and repulsion between hydrophilic groups (in water) can be translated straightforwardly to protein environment.

Hydropathy values of prokaryotic proteins are found to be higher than those of eukaryotes. The lower hydrophobicity of eukaryotic proteins can be atributed to differences in the amino acid composition, such as higher presence of cysteine and corresponding disulfide bonds level of eukaryotic proteins compared to prokaryotic proteins most probably compensates for their lower hydrophobicity. This goes in favor of the importance of hydrophobicity regarding protein stability.

In our dataset, about 75% of our sequences are concentrated in the range from –0.4 to 0.4 and 25% are in the range of low GRAVY, from -0.4 to -0.75, representing the fraction of our proteins with most pronounced hydrophilic properties (Fig 5.6). The average of the GRAVY score for entire group of our proteins were -0.21, that does not differ much from the highly soluble subset of proteins with solubility % over 70, where the average of GRAVY score was -0.28.

Overall hydrophobicity is found to be the most important factor for an ORF to yield a soluble expression product (28), in fact the more negative the GRAVY value becomes, the more likely that an ORF exhibits soluble protein expression. Even though, soluble expression depends on other factors and can not be accurately predicted by bioinformatics methods and our results confirm that slightly negative GRAVY values are one of the properties of the very soluble thermostable proteins in our dataset, but at the same time are not their most prominent or the exclusive property.

Relationship between solubility, thermostability and protein class

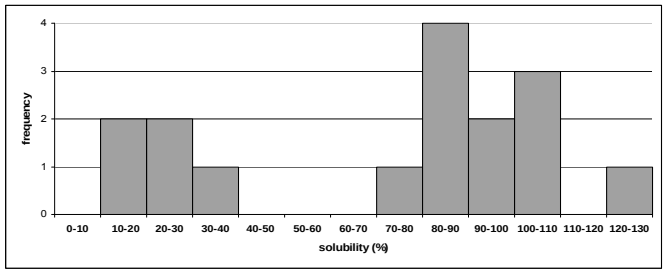
Next, we compared the protein classes of our isolated sequences based on Structural Classification of Proteins database (SCOP) that comprehensively organizes all proteins with known structures based on

their evolutionary and structural relationships (35). In order to address the possible correlation of our investigated thermostability/solubility properties with protein structure, we compared the structures from the SCOP database in respect to our protein grouping. SCOP classification is organized on hierarchical levels: class, fold, superfamily, and family. Within those, superfamilies and families are said to have a common fold if their proteins have the same major secondary structures in the same arrangement with the same topological connections. Folds are assigned to structural classes: all-α (SCOP class a), all-β (class b), α/β (class c), and $\alpha+\beta$ (class d). Within our set of results, the classes are defined at the domain level, and if the protein has more than one domain, it belongs to the class defined by all of the present domains.

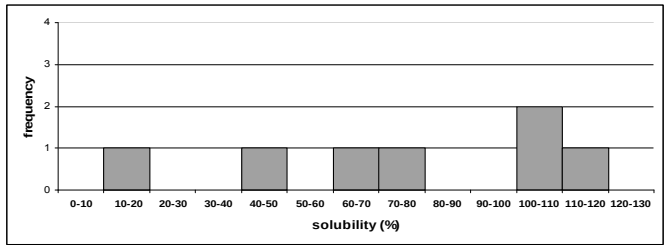
Frequency of certain protein class within our data is compared to the frequency of the same class within the complete genome of *E. coli*. All-α class in our datasets has been statistically substantially more frequent in thermally selected *E. coli* protein subset. It is expected result as this group contains some of the most abundant as well as stable proteins like ferredoxin.

The α-helical ferredoxin domain in *E. coli* contains two Fe4-S4 clusters, typical of bacterial ferredoxin. Iron-sulphur proteins play an important role in electron transfer processes and in various enzymatic reactions. The α-helical ferredoxin domain is present in several proteins involved in redox reactions. Within the class of all-α are also various ribosomal proteins of *E. coli*, while the group of DNA/RNA binding proteins is especially frequent. The soluble proteins also seem to have a more favored helix and strand composition based on the known secondary structural propensities of amino acids, so that evolutionary pressure is directed against protein aggregation. Pronounced increased frequency of alpha proteins in our dataset (Fig. 5.8) is in agreement with our previous results that addressed only thermostability properties, but now we can expand the same conclusions regarding lack of propensity to aggregate.

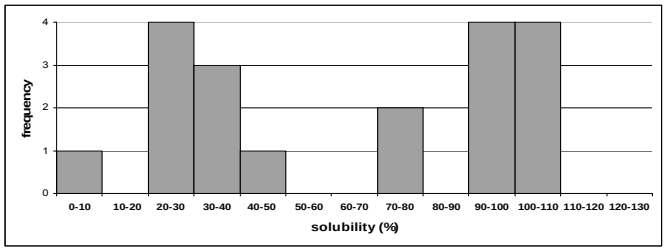
a)



b)



c)



d)

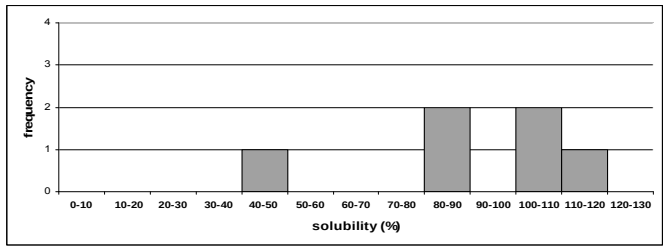


Fig. 5.7 Correlation between solubility and tertiary structure of "*S*" sequences. Histograms of solubility in the SCOP classes: a) all α proteins, b) all β proteins, c) α/β proteins, d) $\alpha+\beta$ proteins

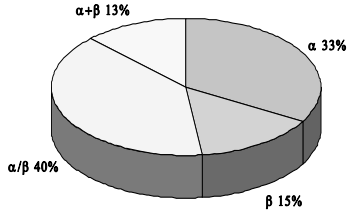


Fig. 5.8. Frequency of protein classes within "S" dataset: α–33%; β–15%; α+β-13%; α/β–40%

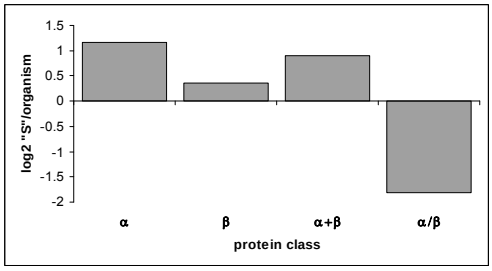


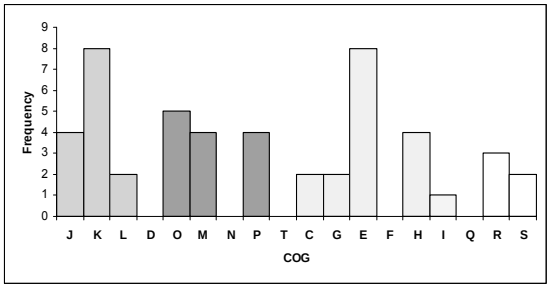
Fig. 5.9. Relative frequency of protein classes within "S" dataset in respect to the entire organism

What is presently hard to explain is elevated frequency of α+β class and strongly diminished frequency of α/β class of protein sequences in our dataset compared to the entire organism (Fig. 5.9). In our dataset of thermally selected *E. coli* subproteome, lower frequency of α/β proteins is very pronounced. Among these proteins are those that provide important metabolic functions in various biosynthetic processes. Class of α+β proteins presented by mainly antiparallel β sheets (segregated alpha and beta regions) is significantly more frequent in respect to the entire organism. Observation that in the *E. coli* cytosol, a fraction of the newly synthesised proteins requires the activity of molecular chaperones for folding to the native state that belong to this SCOP class, is not applicable here, as we are dealing with a subset of our initial data, namely the one that can be expressed without the chaperone presence.

Relationship between protein thermostability, solubility and cellular biological function

We have discussed in previous chapter of this thesis which cellular processes have more of a thermostable character based on the functions of our identified protein sequences that were able to survive thermal treatment. Cellular functions that were discussed are not attributed to the identified sequences according to a particular fold but orthologous evolutionary connections. As presented in the previous chapter, in thermally treated cytosolic extract of *E.coli*, thermally enriched

functional categories (in respect to the complete genome of the organism) were: P – inorganic ion transport and metabolism, K – transcription, H – Coenzyme transport and metabolism, T – signal transduction mechanism, J – Translation, E - aminoacid transport and metabolism, O - chaperones, protein turnover and posttranslational modifications and S – Function unknown (Fig 4.12).



5.10. GOG functional annotation in "*S*" dataset: J,K,L – Information storage and processing; D,O,M,N,P,T – Cellular processes; C,G,E,F,H,I,Q – Metabolism; R,S – Poorly characterized

Functional classification of identified proteins within "S" dataset had representatives in most of the COG categories (Fig. 5.10), but represent higher frequencies of certain categories: within the group Information storage and processing (J,K,L): K – transcription, within the group Metabolism (C,G,E,F,H,I,Q): E - aminoacid transport and metabolism and within the group Cellular processes (D,O,M,N,P,T): O – chaperones. When bimodality of the solubility distribution of the sequences was taken into account, and attention focused just on Sol sequences that have more than 70% solubility, functional categories within Sol sequences were 27% - Information storage and processing; 30% - Cellular processes; 40% - Metabolism, and 3% - poorly characterized (Fig 5.11).

It is certainly tempting to claim that these functional categories are the ones, whose presence and intact functions are the most important for the preservation of life, so consequently the properties as elevated stability and preserved solubility even in extreme thermal conditions. However, as we know, other functional groups do not lack importance in cellular function as well. Other possible explanation may exist in the

frequency of the certain categories *per se* within the cell with evolutionary rational. Consequence of normal, ever happening evolutionary processes can be gene duplication and formation of paralogs that can further continue into couple of outcomes: i) when there is no selective advantage in maintenance of the duplicated gene, it is often reduced to a pseudogene and eventually eliminated from the genome, or ii) duplicates that are retained usually are the ones whose gene products are likely to be advantageous to the organism, e.g. in adaptation and life preservation in new or harsh environmental conditions (36). Three types of speciation can follow duplication event and formation of paralogs: neofunctionalization (both or one duplicates gain new function), subfunctionalization (initial functions are divided among duplicates), and conservation of function(s) functions in both duplicates (37). Those that avoid being deleted, often represent features that may give a competitive advantage to organisms when adapting to environmental conditions, so they dominate in particular categories of functional classification that seem to be associated with processes involving interaction with the environment, particularly handling difficult environmental conditions. A positive response to harsh environmental conditions may be presented either by rapid mutation that would make it advantageous in the challenging conditions, or by increasing the gene product as an answer to higher quantity requirement (38). Therefore, we can come out with double conclusions: within our presented dataset, elevated frequencies of certain categories may indicate that they had the origin in the response to harsh or life threatening conditions, and so developed as subgroup with proven highly thermostable properties, as presented in previous chapter. On the other hand, as our dataset “S” is a result of overlapping our thermally enhanced proteome subset with literature-available much larger group of sequences obtained by cell free and chaperone free translation system, it corroborates conclusions that solubility depend strongly on functions – transcription and translation factors, chaperones, proteases and ribosomal proteins have elevated solubility properties (18).

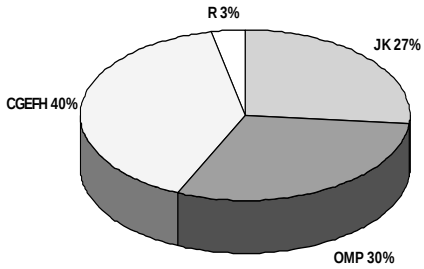


Fig. 5.11 COG functional annotation in Sol subgroup of sequences, more than 70% solubility: J,K – Information storage and processing – 27%; O,M,P – Cellular processes-30%; CGEFH – Metabolism - 40%; R – Poorly characterized - 3%

5.5. Conclusions

Thermally (90^{o}C) selected *E. coli* cytosolic proteome subset was cross-compared with the results published by Taguchi and co-workers regarding bimodal solubility/aggregation distribution of *E. coli* proteome (18) of the proteins expressed in cell-free, chaperone-free system named PURE. The intersection of those two groups of sequences - superstable cytosolic proteins that survived harsh thermal treatment remaining soluble, and chaperone-free-expressed soluble proteins, was investigated in order to check for common determinants between these two important properties: thermostability and solubility. Bimodal pI distribution, prevalence of the proteins below 50 kDa, slightly negative GRAVY values, higher aliphatic index are the properties of the very soluble thermostable proteins that compile, but at the same time are not their most prominent or the exclusive property. The most important observation is that results show maintenance of the bimodal solubility distribution of thermostable protein sequences with obvious increase within soluble subset. Evolutionary cellular guidelines that favor protein thermostability may be common ones that favor solubility of important life preserving sequences.

Our experimental method developed in order to isolate highly thermostable proteome subset has been by many points of view proven to be a good additional tool in mining for proteins with other important properties like lack of propensity to aggregate. Further development of experimental and theoretical as well as computational approaches to

screen for folding and solubility from a thermostable point of view, will facilitate the identification of potential target properties that improve protein solubility and/or modulate misfolding and aggregation. This line of future investigation will not only enable better insight into the processes addressed, but hopefully improve biotechnological processes. Nowadays bacterial cytosol is the major biotechnological factory for recombinant protein production. Therefore, the information regarding factors modulating protein aggregation and thermostability in this specific compartment may be of biotechnological interest. Detailed and specific study of every identified sequence in this research was not a subject to this thesis, but rather the combination of their properties, therefore some of the issues touched here may be interesting continuance and subject to additional investigation.

5.6. References

1. **Minton, A. P.** 2001. The influence of macromolecular crowding and macromolecular confinement on biochemical reactions in physiological media. J Biol Chem **276:**10577-80.
2. **Anfinsen, C. B.** 1973. Principles that govern the folding of protein chains. Science **181:**223-30.
3. **Hartl, F. U., A. Bracher, and M. Hayer-Hartl.** Molecular chaperones in protein folding and proteostasis. Nature **475:**324-32.
4. **Hendrick, J. P., and F. U. Hartl.** 1995. The role of molecular chaperones in protein folding. Faseb J **9:**1559-69.
5. **Ellis, J.** 1987. Proteins as molecular chaperones. Nature **328:**378-9.
6. **Dill, K. A.** 1990. Dominant forces in protein folding. Biochemistry **29:**7133-55.
7. **Agostini, F., M. Vendruscolo, and G. G. Tartaglia.** Sequence-Based Prediction of Protein Solubility. J Mol Biol.
8. **Smialowski, P., A. J. Martin-Galiano, A. Mikolajka, T. Girschick, T. A. Holak, and D. Frishman.** 2007. Protein solubility: sequence based prediction and experimental verification. Bioinformatics **23:**2536-42.
9. **Shen, W., S. Yun, B. Tam, K. Dalal, and F. F. Pio.** 2005. Target selection of soluble protein complexes for structural proteomics

studies. Proteome Sci **3:**3.

10. **Magnan, C. N., A. Randall, and P. Baldi.** 2009. SOLpro: accurate sequence-based prediction of protein solubility. Bioinformatics **25:**2200-7.
11. **Ross, C. A., and M. A. Poirier.** 2004. Protein aggregation and neurodegenerative disease. Nat Med **10 Suppl:**S10-7.
12. **Gomes, C. M., J. B. Vicente, A. Wasserfallen, and M. Teixeira.** 2000. Spectroscopic studies and characterization of a novel electron-transfer chain from Escherichia coli involving a flavorubredoxin and its flavoprotein reductase partner. Biochemistry **39:**16230-7.
13. **Kawarabayasi, Y., Y. Hino, H. Horikawa, K. Jin-no, M. Takahashi, M. Sekine, S. Baba, A. Ankai, H. Kosugi, A. Hosoyama, S. Fukui, Y. Nagai, K. Nishijima, R. Otsuka, H. Nakazawa, M. Takamiya, Y. Kato, T. Yoshizawa, T. Tanaka, Y. Kudoh, J. Yamazaki, N. Kushida, A. Oguchi, K. Aoki, S. Masuda, M. Yanagii, M. Nishimura, A. Yamagishi, T. Oshima, and H. Kikuchi.** 2001. Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, Sulfolobus tokodaii strain7. DNA Res **8:**123-40.
14. **Teixeira, M., R. Batista, A. P. Campos, C. Gomes, J. Mendes, I. Pacheco, S. Anemuller, and W. R. Hagen.** 1995. A seven-iron ferredoxin from the thermoacidophilic archaeon Desulfurolobus ambivalens. Eur J Biochem **227:**322-7.
15. **Li, W. F., X. X. Zhou, and P. Lu.** 2005. Structural features of thermozymes. Biotechnol Adv **23:**271-81.
16. **Prosinecki, V., H. M. Botelho, S. Francese, G. Mastrobuoni, G. Moneti, T. Urich, A. Kletzin, and C. M. Gomes.** 2006. A proteomic approach toward the selection of proteins with enhanced intrinsic conformational stability. J Proteome Res **5:**2720-6.
17. **Cowan, D. A.** 1992. Biotechnology of the Archaea. Trends Biotechnol **10:**315-23.
18. **Niwa, T., B. W. Ying, K. Saito, W. Jin, S. Takada, T. Ueda, and H. Taguchi.** 2009. Bimodal protein solubility distribution revealed by an aggregation analysis of the entire ensemble of Escherichia coli proteins. Proc Natl Acad Sci U S A **106:**4201-6.
19. **Wilkinson, D. L., and R. G. Harrison.** 1991. Predicting the solubility of recombinant proteins in Escherichia coli. Biotechnology (N Y) **9:**443-8.
20. **Christendat, D., A. Yee, A. Dharamsi, Y. Kluger, A. Savchenko, J. R. Cort, V. Booth, C. D. Mackereth, V. Saridakis, I. Ekiel, G.**

Kozlov, K. L. Maxwell, N. Wu, L. P. McIntosh, K. Gehring, M. A. Kennedy, A. R. Davidson, E. F. Pai, M. Gerstein, A. M. Edwards, and C. H. Arrowsmith. 2000. Structural proteomics of an archaeon. Nat Struct Biol **7:**903-9.

21. **Goh, C. S., N. Lan, S. M. Douglas, B. Wu, N. Echols, A. Smith, D. Milburn, G. T. Montelione, H. Zhao, and M. Gerstein.** 2004. Mining the structural genomics pipeline: identification of protein properties that affect high-throughput experimental analysis. J Mol Biol **336:**115-30.
22. **Takemoto, K., T. Niwa, and H. Taguchi.** Difference in the distribution pattern of substrate enzymes in the metabolic network of Escherichia coli, according to chaperonin requirement. BMC Syst Biol **5:**98.
23. **Fujiwara, K., Y. Ishihama, K. Nakahigashi, T. Soga, and H. Taguchi.** A systematic survey of in vivo obligate chaperonin-dependent substrates. Embo J **29:**1552-64.
24. **Shimizu, Y., A. Inoue, Y. Tomari, T. Suzuki, T. Yokogawa, K. Nishikawa, and T. Ueda.** 2001. Cell-free translation reconstituted with purified components. Nat Biotechnol **19:**751-5.
25. **Koschorreck, M., M. Fischer, S. Barth, and J. Pleiss.** 2005. How to find soluble proteins: a comprehensive analysis of alpha/beta hydrolases for recombinant expression in E. coli. BMC Genomics **6:**49.
26. **Cromwell, M. E., E. Hilario, and F. Jacobson.** 2006. Protein aggregation and bioprocessing. Aaps J **8:**E572-9.
27. **Nandi, S., N. Mehra, A. M. Lynn, and A. Bhattacharya.** 2005. Comparison of theoretical proteomes: identification of COGs with conserved and variable pI within the multimodal pI distribution. BMC Genomics **6:**116.
28. **Luan, C. H., S. Qiu, J. B. Finley, M. Carson, R. J. Gray, W. Huang, D. Johnson, J. Tsao, J. Reboul, P. Vaglio, D. E. Hill, M. Vidal, L. J. Delucas, and M. Luo.** 2004. High-throughput expression of C. elegans proteins. Genome Res **14:**2102-10.
29. **Idicula-Thomas, S., and P. V. Balaji.** 2005. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in Escherichia coli. Protein Sci **14:**582-92.
30. **Schwartz, R., C. S. Ting, and J. King.** 2001. Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. Genome Res **11:**703-9.

31. **Kiraga, J., P. Mackiewicz, D. Mackiewicz, M. Kowalczuk, P. Biecek, N. Polak, K. Smolarczyk, M. R. Dudek, and S. Cebrat.** 2007. The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. BMC Genomics **8:**163.
32. **Ikai, A.** 1980. Thermostability and aliphatic index of globular proteins. J Biochem **88:**1895-8.
33. **King, J., C. Haase-Pettingell, A. S. Robinson, M. Speed, and A. Mitraki.** 1996. Thermolabile folding intermediates: inclusion body precursors and chaperonin substrates. Faseb J **10:**57-66.
34. **Kyte, J., and R. F. Doolittle.** 1982. A simple method for displaying the hydropathic character of a protein. J Mol Biol **157:**105-32.
35. **Murzin, A. G., S. E. Brenner, T. Hubbard, and C. Chothia.** 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol **247:**536-40.
36. **Hooper, S. D., and O. G. Berg.** 2003. On the nature of gene innovation: duplication patterns in microbial genomes. Mol Biol Evol **20:**945-54.
37. **Hahn, M. W.** 2009. Distinguishing among evolutionary models for the maintenance of gene duplicates. J Hered **100:**605-17.
38. **Bratlie, M. S., J. Johansen, B. T. Sherman, W. Huang da, R. A. Lempicki, and F. Drablos.** Gene duplications in prokaryotes can be associated with environmental adaptation. BMC Genomics **11:**588.

Chapter 6

CONCLUDING REMARKS

Concluding remarks

This thesis presents a novel strategy which allows mining proteomes for proteins having enhanced stability properties, and further incorporates our results with other approaches that represent strategies to address protein conformation and stability properties at a proteomic scale. Following our interest in the study of conformational properties of thermophilic proteins, we have selected archaeal organism *Sulfurisphaera sp.* from the order *Sulfolobales* as a model organism to implement this methodology: the identification of hyperstable proteins in a thermophilic background required drastic perturbation protocols (up to 4 days at 90°C) which nevertheless resulted in the identification of a subset of proteins which remained folded after the perturbation and provided an insightful perspective into the type of essential cellular processes requiring particularly resistant proteins. In fact, many of the identified proteins are involved in stress response mechanisms that aim at protecting nucleic acids and proteins from aggression elements such as thermal stress and reactive oxygen species. In this respect, this approach highlighted not only proteins interesting for subsequent stability studies, but also on key metabolic processes which may have themselves a "thermophilic character". Another valuable outcome of this study is the applicability of the experimental approach that turned out to be highly efficient in selecting the thermostable soluble subproteome of the organism in question, but can also be applied on any other proteome or complex protein mixture of diverse origins.

Using the established thermal treatment (90^{o}C) and including closely related archaeal hyperthemophilic organism *S. solfataricus* and mesophilic bacterium *E. coli*, we have validated the results regarding the profiling proteins at a proteomic scale according to their enhanced stability. iTRAQ protein sequences identification and quantification of this "thermoproteome" and bioinformatics analysis gave an insight into possible relationship between thermostability on one side and amino acid content, physicochemical properties and biological function on the other. It was found that *per se* presence or absence of certain type of amino acid residues or group is not a sufficient prerequisite for difference in stability, but rather the position and subsequent interaction

of the residues with the surrounding ones can lead towards alteration in stability properties which is in agreement with various other studies. SCOP folds are also not very strongly biased in respect to the thermostability, but it was obvious that all a and all b proteins were favored in both thermophilic and mesophilic organism. As a general observation regarding this results, is that none of the so-called predictive thermostability measures have strong predictive capabilities and the most obvious explanation is that their combination may or may not lead towards enhanced thermal stability, which is influenced by the organisms' ecological surroundings and natural habitat that evolutionary select appropriate stability determinants, that is in accordance with studies so far regarding thermostability basics of proteins and range of factors contributing to it, but no general mechanism that can be pointed out as the most important factor for increased thermostability was found. The observation regarding cellular processes that require especially stable participants is once again confirmed that COG functional category U - intracellular trafficking and secretion, was the most enhanced functional category in *S. solfataricus*.

In order to further incorporate our results with other approaches that represent strategies to address protein stability properties at a proteomic scale, thermally selected *E. coli* cytosolic proteome subset was cross-compared with the results published by Taguchi and co-workers regarding bimodal solubility/aggregation distribution of *E. coli* proteome of the proteins expressed in cell-free, chaperone-free system named PURE. The intersection of those two groups of sequences - superstable cytosolic proteins that survived harsh thermal treatment remaining soluble, and chaperone-free-expressed soluble proteins, was investigated in order to check for common determinants between these two important properties: thermostability and solubility. Bimodal pI distribution, prevalence of the proteins below 50 kDa, slightly negative GRAVY values, higher aliphatic index are the properties of the very soluble thermostable proteins that compile, but at the same time are not their most prominent or the exclusive property. The most important observation is that results show maintenance of the bimodal solubility distribution of thermostable protein sequences with evident increase

within soluble subset. Evolutionary cellular guidelines that favor protein thermostability may be common ones that favor solubility of important life preserving sequences.

Our experimental method developed in order to isolate highly thermostable proteome subset has been by many points of view proven to be a good additional tool in mining for proteins with other important properties like lack of propensity to aggregate. Further development of experimental and theoretical as well as computational approaches to screen for folding and solubility from a thermostable point of view, will facilitate the identification of potential target properties that improve protein solubility and/or modulate misfolding and aggregation. This line of future investigation will enable better insight into a processes addressed, and might hopefully help into designing efficient biological tools to improve biotechnological processes. Nowadays bacterial cytosol is the major biotechnological factory for recombinant protein production. Therefore, the information regarding factors modulating protein aggregation and thermostability in this specific compartment can have further biotechnological interest. Also, as detailed and specific study of every identified sequence in this research was not a subject to this thesis, but rather the combination of their properties, so some of the issues touched here may be interesting continuance and subject to additional investigation. Also, it would certainly be motivating to expand these results including other representative organisms from different natural habitats, with wide range of optimal growth temperatures and provide better insight in evolutionary guidance of selecting proteins with hyper stable properties.

Appendix I

Escherichia coli IDENTIFIED PROTEIN SEQUENCES IN "THERMALLY ENHANCED" GROUP

Appendix I
Data: *Escherichia coli* "Thermally enhanced" group of identified protein sequences – page 1

GI	*pI*	*MW(kDa)*	*Domain*	*Class*	*Fold*	*Superfamily*
16128058	6.46	33,384	46689	α	DNA/RNA-binding 3-helical bundle	Homeodomain-like
16128163	5.22	30,423	46934	α	RuvA C-terminal domain-like	UBA-like
16128425	9.69	9,226	47729	α	IHF-like DNA-binding proteins	IHF-like DNA-binding proteins
16128659	5.68	16,795	46785	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
16128717	8.58	28,231	48452	α	alpha-alpha superhelix	TPR-like
16129198	5.44	15,540	81273	α	H-NS histone-like proteins	H-NS histone-like proteins
16129281	5.74	36,134	47413	α	lambda repressor-like DNA-binding domains	lambda repressor-like DNA-binding domains
16129929	5.31	35,856	46785	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
16130216	5.40	20,538	46548	α	Globin-like	alpha-helical ferredoxin
16130817	6.40	33,472	46785	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
16131061	4.53	54,871	47794	α	SAM domain-like	Rad51 N-terminal domain-like
16131570	8.77	52,551	48295	α	DNA/RNA-binding 3-helical bundle	TrpR-like
16131776	5.40	12,141	47598	α	Ribbon-helix-helix	Ribbon-helix-helix

Appendix I

Data: *Escherichia coli* "Thermally enhanced" group of identified protein sequences – page 2

GI	*pI*	*MW(kDa)*	*Domain*	*Class*	*Fold*	*Superfamily*
16131816	4.60	12,295	48300	α	Ribosomal protein L7/12	
16131968	4.85	57,329	48592	α	GroEL equatorial domain-like	GroEL equatorial domain-like
49176392	5.47	26,079	46785	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
90111554	4.71	17,641	46557	α	Long alpha-hairpin	GreA transcript cleavage protein
16128208	5.97	20,815	53697	α/β	SIS domain	SIS domain
16128371	5.64	28,145	51735	α/β	NAD(P)-binding Rossmann-fold domains	NAD(P)-binding Rossmann-fold domains
16128500	5.23	45,694	53187	α/β	Phosphorylase/hydrolase-like	Zn-dependent exopeptidases
16128638	8.61	33,420	53850	α/β	Periplasmic binding protein-like II	Periplasmic binding protein-like II
16128746	5.56	24,139	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
16128828	6.83	26,829	53850	α/β	Periplasmic binding protein-like II	Periplasmic binding protein-like II
16128831	5.79	26,929	53850	α/β	Periplasmic binding protein-like II	Periplasmic binding protein-like II
16128844	9.35	63,589	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
16128975	5.84	28,898	53474	α/β	alpha/beta-Hydrolases	alpha/beta-Hydrolases

Appendix I

Data: *Escherichia coli* "Thermally enhanced" group of identified protein sequences – page 3

GI	*pI*	*MW(kDa)*	*Domain*	*Class*	*Fold*	*Superfamily*
16128986	9.03	50,766	53448	α/β	Nucleotide-diphospho-sugar transferases	Nucleotide-diphospho-sugar transferases
16128997	5.53	26,890	89550	α/β	7-stranded beta/alpha barrel	PHP domain-like
16129281	5.74	36,134	53822	α/β	Periplasmic binding protein-like I	Periplasmic binding protein-like I
16129619	5.69	43,909	53335	α/β	S-adenosyl-L-methionine-dependent methyltransferases	S-adenosyl-L-methionine-dependent methyltransferases
16129929	5.31	35,856	53850	α/β	Periplasmic binding protein-like II	Periplasmic binding protein-like II
16129961	5.06	46,110	53720	α/β	ALDH-like	ALDH-like
16130174	5.38	40,843	51695	α/β	TIM beta/alpha-barrel	PLC-like phosphodiesterases
16130245	5.62	27,991	53850	α/β	Periplasmic binding protein-like II	Periplasmic binding protein-like II
16130386	5.18	17,659	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
16130420	5.89	13,399	52833	α/β	Thioredoxin fold	Thioredoxin-like
16130451	4.98	65,652	53067	α/β	Ribonuclease H-like motif	Actin-like ATPase domain
16130817	6.40	33,472	53850	α/β	Periplasmic binding protein-like II	Periplasmic binding protein-like II
16131570	8.77	52,551	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases

Appendix I

Data: *Escherichia coli* "Thermally enhanced" group of identified protein sequences – page 4

GI	*pI*	*MW(kDa)*	*Domain*	*Class*	*Fold*	*Superfamily*
16131823	5.54	23,015	51391	α/β	TIM beta/alpha-barrel	Thiamin phosphate synthase
16131836	5.53	57,329	52335	α/β	Methylglyoxal synthase-like	Methylglyoxal synthase-like
16131836	5.53	57,329	64197	α/β	Cytidine deaminase-like	AICAR transformylase domain of bifunctional purine biosynthesis enzyme ATIC
16131968	4.85	57,329	52029	α/β	The "swivelling" beta/beta/alpha domain	GroEL apical domain-like
49176430	6.31	40,876	53901	α/β	Thiolase-like	Thiolase-like
90111448	5.32	22,533	53271	α/β	PRTase-like	PRTase-like
145698255	7.89	149,027	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
16128163	5.22	30,423	54713	α+β	Elongation factor Ts (EF-Ts)	
16128165	6.44	20,638	55194	α+β	RRF/tRNA synthetase additional domain-like	Ribosome recycling factor
16128500	5.23	45,694	55031	α+β	Ferredoxin-like	Bacterial exopeptidase dimerisation domain
16129441	5.57	15,088	82784	α+β	OsmC-like	OsmC-like
16129756	5.45	42,561	55961	α+β	TBP-like	Bet v1-like
16129817	4.71	26,422	75625	α+β	YebC-like	YebC-like

Appendix I

Data: *Escherichia coli* "Thermally enhanced" group of identified protein sequences – page 5

GI	*pI*	*MW(kDa)*	*Domain*	*Class*	*Fold*	*Superfamily*
16130091	6.80	24,830	55620	α+β	T-fold	Tetrahydrobiopterin biosynthesis enzymes-like
16130341	5.65	9,119	55594	α+β	HPr-like	HPr-like
16130937	4.90	23,353	55821	α+β	YrdC/RibB	YrdC/RibB
16131061	4.53	54,871	69705	α+β	Transcription factor NusA	
16131061	4.53	54,871	54814	α+β	Alpha-lytic protease prodomain-like	Prokaryotic type KH domain (KH-domain type II)
16131266	5.69	46,162	110997	α+β	Ferredoxin-like	Sporulation related repeat (SPOR
16131816	4.60	12,295	54736	α+β	ClpS-like	ClpS-like
16131968	4.85	57,329	54849	α+β	GroEL-intermediate domain like	GroEL-intermediate domain like
16131984	6.55	123,967	82689	α+β	Ferredoxin-like	Mechanosensitive channel protein MscS (YggB)
90111554	4.71	17,641	54534	α+β	FKBP-like	FKBP-like
90111586	4.35	32,534	64397	α+β	Hsp33 domain	Hsp33 domain
16128058	6.46	33,384	51215	β	Double-stranded beta-helix	Regulatory protein AraC
16128982	4.98	41,137	49503	β	Cupredoxin-like	Cupredoxins

Appendix I
Data: *Escherichia coli* "Thermally enhanced" group of identified protein sequences – page 6

GI	***pI***	***MW(kDa)***	***Domain***	***Class***	***Fold***	***Superfamily***
16129019	5.57	20,912	101874	β	Streptavidin-like	Hypothetical protein TT1927B
16129756	5.45	42,561	50022	β	ISP domain	ISP domain
16130451	4.98	65,652	100920	β	Heat shock protein 70kD (HSP70)	
16131061	4.53	54,871	50249	β	OB-fold	Nucleic acid-binding proteins
16131188	10.21	11,316	50104	β	SH3-like barrel	Translation proteins SH3-like domain
16131290	4.52	20,998	89360	β	HesB-like domain	HesB-like domain
16131984	6.55	123,967	50182	β	Sm-like fold	Sm-like ribonucleoproteins
145698236	6.28	22,497	89392	β	Lipoprotein localization factors LolAB	Lipoprotein localization factors LolAB

NCBI **GI** sequence identification number: http://www.ncbi.nlm.nih.gov/protein/
pI Theoretical isoelectic point http://web.expasy.org/protparam/
MW Molecular mass http://web.expasy.org/protparam/
SCOP **Domain**, **Class**, **Fold**, **Superfamily:** http://supfam.cs.bris.ac.uk/SUPERFAMILY/cgi-bin/search.cgi

Appendix II

Sulfolobus solfataricus IDENTIFIED PROTEIN SEQUENCES IN "THERMALLY ENHANCED" GROUP

Appendix II

Data: *Sulfolobus solfataricus* "Thermally enhanced" group of identified protein sequences – page 1

GI	*pI*	*MW(kDa)*	*Domain*	*Class*	*Fold*	*Superfamily*
15897061	8.43	54,101	46785	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
15897552	6.11	51,795	48557	α	L-aspartase-like	L-aspartase-like
15897632	6.38	14,543	46579	α	Long alpha-hairpin	Prefoldin
15897850	9.16	50,047	47364	α	Four-helical up-and-down bundle	Domain of the SRP/SRP receptor G-proteins
15897850	9.16	50,047	47446	α	Signal peptide-binding domain	Signal peptide-binding domain
15897861	7.49	19,392	101424	α	alpha-alpha superhelix	Hypothetical protein ST1625
15897970	7.72	14,071	46785	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
15898746	5.95	75,724	48208	α	alpha/alpha toroid	Six-hairpin glycosidases
15898968	5.24	30,888	48452	α	alpha-alpha superhelix	TPR-like
15899115	8.84	36,792	46548	α	Globin-like	alpha-helical ferredoxin
15899217	9.71	11,072	46785	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
15899325	5.54	26,668	48613	α	Heme oxygenase-like	Heme oxygenase-like
15899368	5.44	16,034	47240	α	Ferritin-like	Ferritin-like

Appendix II
Data: *Sulfolobus solfataricus* "Thermally enhanced" group of identified protein sequences– page 2

GI	*pI*	*MW(kDa)*	*Domain*	*Class*	*Fold*	*Superfamily*
15899376	9.25	18,934	46785	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
15896989	7.29	83,674	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
15897113	9.10	31,017	64005	α/β	Undecaprenyl diphosphate synthase	Undecaprenyl diphosphate synthase
15897401	9.62	50,817	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
15897836	5.78	52,839	75304	α/β	Amidase signature (AS) enzymes	Amidase signature (AS) enzymes
15897844	9.22	108,518	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
15897850	9.16	50,047	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
15898001	7.73	47,778	53686	α/β	Tryptophan synthase beta subunit-like PLP-dependent enzymes	Tryptophan synthase beta subunit-like PLP-dependent enzymes
15898070	6.12	54,604	53720	α/β	ALDH-like	ALDH-like
15898104	5.39	20,623	102405	α/β	Putative lysine decarboxylase	Putative lysine decarboxylase
15898119	9.27	36,913	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
15898174	8.65	69,867	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
15898288	9.46	15,906	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases

Appendix II
Data: *Sulfolobus solfataricus* "Thermally enhanced" group of identified protein sequences– page 3

GI	*pI*	*MW(kDa)*	*Domain*	*Class*	*Fold*	*Superfamily*
15898485	6.37	57,620	53067	α/β	Ribonuclease H-like motif	Actin-like ATPase domain
15898822	7.12	60,110	53795	α/β	PEP carboxykinase-like	PEP carboxykinase-like
15898822	7.12	60,110	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
15898872	4.93	67,086	52743	α/β	Subtilisin-like	Subtilisin-like
15898877	5.35	64,370	51445	α/β	TIM beta/alpha-barrel	(Trans)glycosidases
15898960	4.98	117,475	52743	α/β	Subtilisin-like	Subtilisin-like
15899094	4.84	23,490	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
15899341	9.17	36,331	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
15899357	5.51	14,042	75169	α/β	YchN-like	YchN-like
15899432	8.10	35,832	51735	α/β	NAD(P)-binding Rossmann-fold domains	NAD(P)-binding Rossmann-fold domains
15899531	6.37	70,228	53323	α/β	Pyruvate-ferredoxin oxidoreductase	
15899531	6.37	70,228	52518	α/β	Thiamin diphosphate-binding fold (THDP-binding)	Thiamin diphosphate-binding fold (THDP-binding)
15899531	6.37	70,228	52922	α/β	TK C-terminal domain-like	TK C-terminal domain-like

Appendix II

Data: *Sulfolobus solfataricus* "Thermally enhanced" group of identified protein sequences– page 4

GI	*pI*	*MW(kDa)*	*Domain*	*Class*	*Fold*	*Superfamily*
15899565	7.83	39,113	52540	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
15899608	8.43	76,330	56784	α/β	HAD-like	HAD-like
15899726	6.04	56,691	51445	α/β	TIM beta/alpha-barrel	(Trans)glycosidases
15899847	8.14	42,074	89796	α/β	CoA-transferase family III (CaiB/BaiF)	CoA-transferase family III (CaiB/BaiF)
15899922	6.24	28,509	53448	α/β	Nucleotide-diphospho-sugar transferases	Nucleotide-diphospho-sugar transferases
15897061	8.43	54,101	55681	α+β	Class II aaRS and biotin synthetases	Class II aaRS and biotin synthetases
15897128	4.37	10,136	54984	α+β	Ferredoxin-like	eEF-1beta-like
15897250	6.65	30,128	54211	α+β	Ribosomal protein S5 domain 2-like	Ribosomal protein S5 domain 2-like
15897250	6.65	30,128	55060	α+β	Ferredoxin-like	GHMP Kinase
15897326	8.65	10,351	55120	α+β	Pseudouridine synthase	Pseudouridine synthase
15898872	4.93	67,086	54897	α+β	Ferredoxin-like	Protease propeptides/inhibitors
15898960	4.98	117,475	54897	α+β	Ferredoxin-like	Protease propeptides/inhibitors
15899115	8.84	36,792	54292	α+β	beta-Grasp (ubiquitin-like)	2Fe-2S ferredoxin-like

Appendix II

Data: *Sulfolobus solfataricus* "Thermally enhanced" group of identified protein sequences – page 5

GI	***pI***	***MW(kDa)***	***Domain***	***Class***	***Fold***	***Superfamily***
15899165	5.81	15,972	55620	α+β	T-fold	Tetrahydrobiopterin biosynthesis enzymes-like
15899306	5.06	11,044	54862	α+β	Ferredoxin-like	4Fe-4S ferredoxins
15899317	7.74	15,582	54631	α+β	CBS-domain	CBS-domain
15899355	6.23	8,400	64307	α+β	IF3-like	SirA-like
15899603	5.37	29,258	56281	α+β	Metallo-hydrolase/oxidoreductase	Metallo-hydrolase/oxidoreductase
15899608	8.43	76,330	55008	α+β	Ferredoxin-like	HMA
15897929	9.00	16,299	51230	β	Barrel-sandwich hybrid	Single hybrid motif
15897954	4.94	23,232	49879	β	SMAD/FHA domain	SMAD/FHA domain
15897967	5.95	16,336	51230	β	Barrel-sandwich hybrid	Single hybrid motif
15898554	5.55	25,160	51219	β	Double-stranded beta-helix	TRAP-like
15898877	5.35	64,370	81296	β	Immunoglobulin-like beta-sandwich	E set domains
15898877	5.35	64,370	51011	β	Glycosyl hydrolase domain	Glycosyl hydrolase domain
15899143	4.95	19,694	50324	β	OB-fold	Inorganic pyrophosphatase

Appendix II
Data: *Sulfolobus solfataricus* "Thermally enhanced" group of identified protein sequences – page 6

GI	*pI*	*MW(kDa)*	*Domain*	*Class*	*Fold*	*Superfamily*
15899176	5.31	20,097	49764	β	HSP20-like chaperones	HSP20-like chaperones
15899432	8.10	35,832	50129	β	GroES-like	GroES-like
15899565	7.83	39,113	50331	β	OB-fold	MOP-like
15899608	8.43	76,330	81653	β	Double-stranded beta-helix	Calcium ATPase
15899747	6.47	32,805	63829	β	6-bladed beta-propeller	Calcium-dependent phosphotriesterase
15899863	5.14	9,059	51230	β	Barrel-sandwich hybrid	Single hybrid motif

NCBI **GI** sequence identification number: http://www.ncbi.nlm.nih.gov/protein/
pI Theoretical isoelectic point http://web.expasy.org/protparam/
MW Molecular mass http://web.expasy.org/protparam/
SCOP **Domain**, **Class**, **Fold**, **Superfamily:** http://supfam.cs.bris.ac.uk/SUPERFAMILY/cgi-bin/search.cgi

Appendix III

Escherichia coli IDENTIFIED PROTEIN SEQUENCES IN "S" GROUP

Appendix III

Data: *Escherichia coli*, "S" group of identified protein sequences – page 1

GI	*pI*	*MW (kDa)*	*Domain*	*Solubility (%)*	*Aliphatic index*	*GRAVY*	*Class*	*Fold*	*Superfamily*
16128058	6.46	33,384	46689	27.75	84.86	-0.32	α	DNA/RNA-binding 3-helical bundle	Homeodomain-like
16128163	5.22	30,423	46934	87.2	91.73	-0.12	α	RuvA C-terminal domain-like	UBA-like
16128659	5.68	16,795	46785	109.55	88.31	-0.56	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
16128717	8.58	28,231	48452	19.59	76.58	-0.47	α	alpha-alpha superhelix	TPR-like
16129198	5.44	15,540	81273	103.34	82.7	-0.75	α	H-NS histone-like proteins	H-NS histone-like proteins
16131968	4.85	57,329	48592	103.34	96.86	-0.01	α	GroEL equatorial domain-like	GroEL equatorial domain-like
16129281	5.74	36,134	47413	85.04	96.69	-0.09	α	lambda repressor-like DNA-binding domains	lambda repressor-like DNA-binding domains
16129929	5.31	35,856	46785	20.58	106.49	-0.20	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
16130216	5.40	20,538	46548	72.28	75.33	-0.33	α	Globin-like	alpha-helical ferredoxin
16130817	6.40	33,472	46785	17.03	99.83	-0.17	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
16131061	4.53	54,871	47794	99.46	97.96	-0.28	α	SAM domain-like	Rad51 N-terminal domain-like

Appendix III

Data: *Escherichia coli* ,"S" group of identified protein sequences – page 2

GI	pI	MW (kDa)	Domain	Solubility (%)	Aliphatic index	GRAVY	Class	Fold	Superfamily
90111554	4.71	17,641	46557	87.45	96.27	-0.37	α	Long alpha-hairpin	GreA transcript cleavage protein
16131570	8.77	52,551	48295	38.43	88.63	-0.36	α	DNA/RNA-binding 3-helical bundle	TrpR-like
16131776	5.40	12,141	47598	88.96	78.1	-0.75	α	Ribbon-helix-helix	Ribbon-helix-helix
16131816	4.60	12,295	48300	98.51	106.61	0.30	α	Ribosomal protein L7/12	
49176392	5.47	26,079	46785	129.31	91.7	-0.36	α	DNA/RNA-binding 3-helical bundle	Winged helix DNA-binding domain
16128208	5.97	20,815	53697	91.86	96.61	-0.06	α/β	SIS domain	SIS domain
16128371	5.64	28,145	51735	100.28	97.62	0.32	α/β	NAD(P)-binding Rossmann-fold domains	NAD(P)-binding Rossmann-fold domains
16128500	5.23	45,694	53187	25.43	82.34	-0.23	α/β	Phosphorylase/ hydrolase-like	Zn-dependent exopeptidases
16128746	5.56	24,139	52540	27.11	92.44	0.02	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
16128828	6.83	26,829	53850	73.37	79.55	-0.28	α/β	Periplasmic binding protein-like II	Periplasmic binding protein-like II

Appendix III

Data: *Escherichia coli,* "S" group of identified protein sequences – page 3

GI	*pI*	*MW (kDa)*	*Domain*	*Solubility (%)*	*Aliphatic index*	*GRAVY*	*Class*	*Fold*	*Superfamily*
16128831	5.79	26,929	53850	76.54	80.29	-0.33	α/β	Periplasmic binding protein-like II	Periplasmic binding protein-like II
16128844	9.35	63,589	52540	2.71	96.29	-0.33	α/β	P-loop containing nucleoside triphosphate hydrolases	P-loop containing nucleoside triphosphate hydrolases
16128975	5.84	28,898	53474	17.19	102.82	0.09	α/β	alpha/beta-Hydrolases	alpha/beta-Hydrolases
16128986	9.03	50,766	53448	17.65	106.83	0.30	α/β	Nucleotide-diphospho-sugar transferases	Nucleotide-diphospho-sugar transferases
16128997	5.53	26,890	89550	100.32	88.08	-0.12	α/β	7-stranded beta/alpha barrel	PHP domain-like
16129619	5.69	43,909	53335	26.75	80.94	-0.31	α/β	S-adenosyl-L-methionine-dependent methyltransferases	S-adenosyl-L-methionine-dependent methyltransferases
16129961	5.06	46,110	53720	90.65	93.39	-0.01	α/β	ALDH-like	ALDH-like
16130174	5.38	40,843	51695	39.37	75.2	-0.55	α/β	TIM beta/alpha-barrel	PLC-like phosphodiesterases
16130420	5.89	13,399	52833	100.31	104.87	-0.36	α/β	Thioredoxin fold	Thioredoxin-like

Appendix III

Data: *Escherichia coli,* "S" group of identified protein sequences – page 4

GI	*pI*	*MW (kDa)*	*Domain*	*Solubility (%)*	*Aliphatic index*	*GRAVY*	*Class*	*Fold*	*Superfamily*
90111448	5.32	22,533	53271	108.17	110.58	0.07	α/β	PRTase-like	PRTase-like
16131823	5.54	23,015	51391	10.38	107.35	0.03	α/β	TIM beta/alpha-barrel	Thiamin phosphate synthase
16131968	4.85	57,329	52029	94.55	96.86	-0.01	α/β	The "swivelling" beta/beta/alpha domain	GroEL apical domain-like
16128638	8.61	33,420	53850	93.99	79.83	-0.47	α/β	Periplasmic binding protein-like II	Periplasmic binding protein-like II
49176430	6.31	40,876	53901	10.44	92.14	0.13	α/β	Thiolase-like	Thiolase-like
16128165	6.44	20,638	55194	109.54	90.7	-0.53	α+β	RRF/tRNA synthetase additional domain-like	Ribosome recycling factor
16129441	5.57	15,088	82784	100.96	81.96	-0.12	α+β	OsmC-like	OsmC-like
16130091	6.80	24,830	55620	49.73	98.83	-0.12	α+β	Tetrahydrobiopterin T-fold	Tetrahydrobiopterin biosynthesis enzymes-like
16130341	5.65	9,119	55594	88.37	84.94	-0.17	α+β	HPr-like	HPr-like
16130937	4.90	23,353	55821	116.29	82.35	-0.15	α+β	YrdC/RibB	YrdC/RibB
90111586	4.35	32,534	64397	88.87	90.48	-0.31	α+β	Hsp33 domain	Hsp33 domain
145698236	6.28	22,497	89392	70.61	69.21	-0.50	β	Lipoprotein localization factors LolAB	Lipoprotein localization factors LolAB
16128982	4.98	41,137	49503	42.79	86.75	-0.32	β	Cupredoxin-like	Cupredoxins

Appendix III

Data: *Escherichia coli,* "S" group of identified protein sequences – page 5

GI	*pI*	*MW (kDa)*	*Domain*	*Solubility (%)*	*Aliphatic index*	*GRAVY*	*Class*	*Fold*	*Superfamily*
16129019	5.57	20,912	101874	69.38	76.13	-0.42	β	Streptavidin-like	Hypothetical protein TT1927B
16130451	4.98	65,652	100920	101.91	102.22	-0.03	β	Heat shock protein 70kD (HSP70)	
16131188	10.21	11,316	50104	111.8	89.9	-0.38	β	SH3-like barrel	Translation proteins SH3-like domain
16131984	6.55	123,967	50182	15.57	111.36	0.02	β	Sm-like fold	Sm-like ribonucleoproteins
16129756	5.45	42,561	50022	107.68	73.02	-0.39	β	ISP domain	ISP domain

NCBI **GI** sequence identification number: http://www.ncbi.nlm.nih.gov/protein/

pI Theoretical isoelectic point http://web.expasy.org/protparam/

MW Molecular mass http://web.expasy.org/protparam/

Aliphatic Index, GRAVY (Grand average of hydropathicity) http://web.expasy.org/protparam/

SCOP **Domain**, **Class**, **Fold**, **Superfamily:** http://supfam.cs.bris.ac.uk/SUPERFAMILY/cgi-bin/search.cgi

Solubility (%) http://www.taguchi.bio.titech.ac.jp/eng/paper-e/assets/2009_PNAS_ecoli_proteins_solubility.xls