Genome Sequence Analysis and Characterization of Recombinant Enzymes from the Thermoacidophilic Archaeon *Picrophilus torridus*

Dissertation
Zur Erhaltung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultäten
der Georg-August-Universität zu Göttingen

vorgelegt von
Angel Stoyanov Angelov
aus Botevgrad, Bulgarien

Göttingen 2004

D 7

Referent: Prof. Dr. W. Liebl

Korreferent: Prof. Dr. G. Gottschalk

Tag der mündlichen Prüfung: 29.06.2004

Table of contents

Contents I

Abbreviations V

A. Introduction 1

B. Materials and Methods 6

1. Bacterial strains and growth conditions 6
1.1. Strains and plasmids 6
1.2. Growth media 9
1.3. Growth conditions 11
1.4. Monitoring of growth, culture harvesting and cell fractionation 12
1.5. Storage of strains and control of purity 12

2. DNA manipulations 13
2.1. General techniques 13
2.1.1. DNA purification 13
2.1.2. DNA analysis using agarose gel electrophoresis 15
2.1.3. DNA quantification 16
2.2. DNA isolation 16
2.2.1. Plasmid isolation from *E. coli* 16
2.2.2. Isolation of genomic DNA from *P. torridus* 17
2.3. Enzymatic modification of DNA 18
2.3.1. Restriction 18
2.3.2. Dephosphorilation of linearised DNA 19
2.3.3. Ligation 19
2.4. *In vitro* DNA amplification. Polymerase chain reaction (PCR) 20
2.4.1. Analytical PCR 21
2.4.2. Preparative PCR 22
2.4.3. Colony PCR 23
2.4.4. PCR purification and cloning 23
2.5. Transformation 24
2.5.1. Transformation of *E. coli* 24
2.5.2. Transformation of *S. solfataricus* 24
2.5.3. Transformation of *S. cerevisiae* 25
2.6. Methods used in genome sequencing, assembly and sequence analysis 26

2.6.1. Generation of a whole genome shotgun library 26

2.6.2. Genome assembly and closing of gaps 26

2.6.3. Sequence analysis and annotation 27

3. Protein manipulations and biochemical methods 28

3.1. Determination of protein concentrations 28

3.2. Polyacrilamide gel electrophoresis (PAGE) 28

3.2.1. SDS-PAGE 28

3.2.2. Native PAGE 30

3.3. Concentration and dialysis of proteins 31

3.4. Refolding of proteins 32

3.5. Purification of proteins 34

3.5.1. Heat treatment of crude cellular extracts 34

3.5.2. Fast Protein Liquid Chromatography (FPLC) 34

3.5.3. Purification of maltose binding protein (MBP) fusions 35

3.6. Determination of enzyme activity and biochemical characterisation of enzymes 36

3.6.1. Glucose dehydrogenase (GdhA) enzyme activity measurements 36

3.6.2. Alpha-glucosidase (MalP) enzyme activity measurements 37

3.6.3 Thin layer chromatography (TLC) analysis of the α-glucosidase reaction products 38

C. Results 39

1. Sequencing, assembly and sequence analysis of the *P. torridus* genome 39

1.1. Sequencing, assembly and editing 39

1.2. Gap closure and further editing 40

1.3. Sequence analysis 42

1.3.1. General features of the *P. torridus* genome 42

1.3.2. Comparative analysis of the amino acid composition and the isoelectric point distribution of *P. torridus* proteins 45

1.3.2. Origin of replication 46

1.3.3. Replication, repair and restriction 47

1.3.4. Amino acid metabolism 50

1.3.5. Protein and peptide degradation 51

1.3.6. Protein synthesis and export 51

1.3.7. Transporters 52

1.3.8. Energy metabolism 54
1.3.9. Porphyrin metabolism 58
1.3.10. Oxigen stress genes 59

2. Heterologous expression of *P. torridus* genes 60
2.1. Refolding of solubilised inclusion bodies 61
2.1.1. Refolding by dialysis or rapid dilution 63
2.1.2. Refolding using the *Vectrase* kit 63
2.1.3. Refolding using size exclusion chromatography (SEC) 64
2.2. Use of a weak promotor 66
2.3. Expression of fusion proteins 67
2.3.1. Maltose binding protein tag – pMAL-c2x 67
2.3.2 NusA tag – pET43.1 69
2.4. Other expression systems – the SSV1 virus and *S. solfataricus* 70
2.5. Expression in *S. cerevisiae* 72

3. Purification and biochemical characterisation of glucose-1-dehydrogenase and α-glucosidase from *P. torridus* 73
3.1. *P. torridus* α-glucosidase (MalP) 73
3.1.1. Analysis of the amino acid sequence of MalP 73
3.1.2. Cloning and expression of the *P. torridus mal*P ORF and purification of recombinant MalP 74
3.1.3. Temperature, pH optimum and reaction products of MalP 75
3.2. *P. torridus* glucose-1-dehydrogenase (GdhA) 77
3.2.1. Analysis of the amino acid sequence of GdhA 77
3.2.2. Cloning and expression of the *P. torridus gdh*A ORF and purification of recombinant GdhA 78
3.2.3. Characterisation of recombinant *P. torridus* GdhA 80
3.2.3.1. Substrate specificity and effect of metabolites on the enzyme activity 80
3.2.3.2. Kinetic properties of GdhA 81
3.2.3.3. Temperature, pH optimum and thermoinactivation kinetics 82
3.2.3.4. Influence of Zn^{2+} on the temperature and pH stability of the enzyme 83
3.2.3.5. Stability of GdhA in organic solvents 85
3.2.3.6. Non-enzymatic NADPH degradation 85
3.2.3.7. Identification of the native glucose dehydrogenase activity in *P. torridus* 87

D. Discussion **88**

1. The genome sequence of *P. torridus* **88**
1.1. Sequencing and assembly 88
1.2. Sequence analysis 89
1.2.1. Amino acid composition and isoelectric point distribution 90
1.2.2. Phylogenetic analyses 90
1.3. Bioenergetics and central metabolism 93
1.3.1. The cytoplasm membrane and the generation of membrane potential 93
1.3.1.1. Cell wall and cytoplasm membrane 94
1.3.1.2. Respiration 95
1.3.2. Transporters 96
1.3.3. Carbohydrate metabolism 97
1.3.4. Overview 100

2. Heterologous expression of *P. torridus* genes **102**
2.1. Codon usage 102
2.2. Inclusion body formation 102
2.3. Overexpression of *P. torridus* GdhA and MalP 106

3. Properties of GdhA and MalP **107**
3.1. *P. torridus* α-glucosidase (MalP) 107
3.2. *P. torridus* glucose dehydrogenase(GdhA) 108
3.2.1. Structural basis for stability 109
3.2.2. Native glucose/galactose dehydrogenase activity in P. torridus 111

4. Physiological role of GdhA and the Entner-Doudoroff pathway **111**

E. Summary **116**
F. References **118**
Appendix A:Keto-excess and GC-skew plots **127**
Appendix B: Codon usage table of *P. torridus* **129**
Appendix C: Plasmid maps **131**
Appendix D: *P. torridus* ORFs with predicted signal peptides **136**

Abbreviations

AA	amino acid
AP	alkaline phosphatase
Ap	ampicillin
Ap^r	ampicillin resistance
APS	ammonium persulfate
Asp	aspartic acid
ATP/ADP	adenosine 5’-triphosphate /Adenosine 5’-diphosphate
BCIP	5-Bromo-4-chloro-3-indolyl-phosphate
bp	base pair
BSA	bovine serum albumin
Cm	chloramphenicol
Da	Dalton
dd H_2O	bi-distillated water
DHAP	dihydroxyacetone phosphate
DMF	N,N-dimethyl formamid
DNase	deoxyribonuclease
dNTP	deoxynucleosidetriphosphate
DSMZ	German Collection of Microorganisms and Cell Cultures
E. coli	*Echerichia coli*
EC	Enzyme Commission
ED	Entner-Doudoroff pathway
EDTA	ethylene di-amine tetra-acetic acid
EMP	Embden-Meyerhoff-Parnas pathway
EtOH	ethanol
FBP	fructose-1,6-bisphosphatase
GAP	glyceraldehyde-3-phosphate
GDH	glucose-1-dehydrogenase
GndHCl	guanidine hydrochloride
h	hour
IPTG	Isopropyl β-D-1-thiogalactopyranosid
kb	kilobase pair
KDG	2-keto 3-deoxygluconate
Km	kanamycin
K_m	Michaelis constant
Km^R	kanamycin resistant
LB	Luria-Bertani Broth
MBP	maltose binding protein
MCS	multiple cloning site

min	minutes
M_r	relative molecular weight
NAD(H)	β-Nicotinamide adenine dinucleotide (reduced)
NADP(H)	β-Nicotinamide adenine dinucleotide phosphate (reduced)
NBT	Nitrotetrazolium Blue Chloride
OD	Optical Density
ON	overnight
ORF	open reading frame
OriR	origin of replication
PAGE	polyacrylamide gel electrophoresis
PCR	polymerase chain reaction
PEP	phospho enol pyruvate
PFK	phosphofructokinase
P_i	inorganic phosphate
pI	isoelectric point
PMS	5- methyl phenazonium methyl sulphate
PPi	pyrophosphate anion ($P_2O_7^{4-}$)
PPP	pentose phosphate pathway
PPS	phosphoenolpyruvate synthase
PYK	pyruvate kinase
RNase	ribonuclease
rpm	rounds per minute
RT	room temperature (ca. 23 °C)
s	seconds
SDS	sodium dodecylsulfate
SEC	size exclusion chromatography
TAE	Tris-Acetate-EDTA buffer
TCA	tricarboxylic acid cycle
TE	Tris/EDTA buffer
TLC	thin layer chromatography
Tris	tris-hydroxymethyl-aminomethane
U	Unit (unit of enzyme activity)
UV	ultraviolet (radiation)
v/v	volume per volume
w/v	weight per volume
X-Gal	5-Bromo-4-chloro-3-indolyl-β-D- galactopyranosid

A. Introduction

In the last few decades, the number of species found to inhabit extreme environments has grown rapidly. Additionally, places previously thought to be incompatible with life due to their harsh conditions have been found to be populated. Microorganisms have been found in extremely acidic or alkaline environments (acidophiles or alkaliphiles), such with high salt concentrations (halophiles), under extremely high pressure (piezophiles) and at very low or high temperatures (psychrophiles or hyperthermophiles). It is not uncommon to find in nature a combination of these extremes where microorganisms live, i.e. high temperature and acidity, for example. The organisms, referred to as "extreme acidophiles" are those that have a pH optimum for growth at or below pH 3.0 (Norris and Johnson, 1998). This definition excludes a lot of fungal and yeast species which, despite being able to withstand extreme acidity, have a pH optima of growth near neutrality. Although extremophiles include also different representatives of the *Bacteria* and *Eukarya*, the most abundant organisms found in such places belong to the *Archaea*. The definition of archaea as a separate organismal domain was originally based on 16S RNA sequence analysis (Woese *et al.*; 1990, Fig 1A).

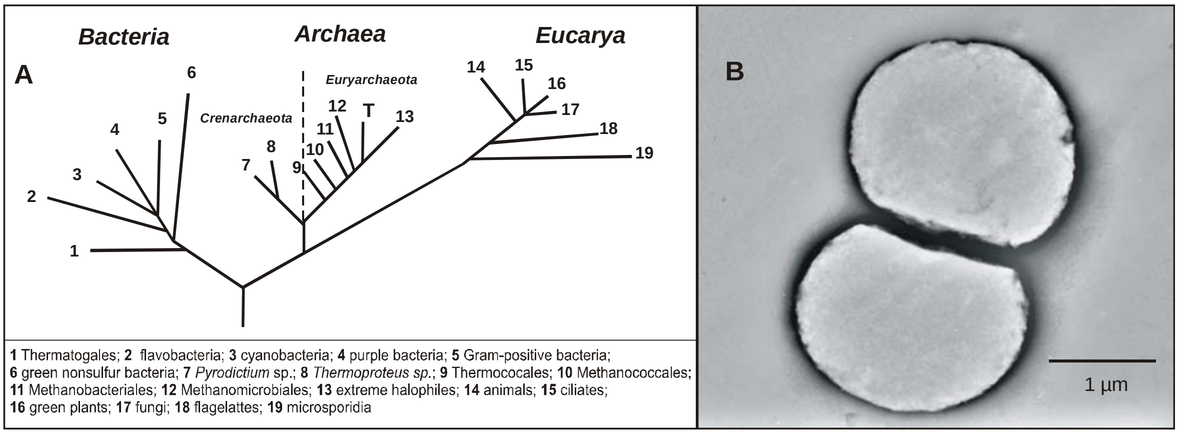

Bacteria
Archaea
Eucarya
A
B
Euryarchaeota
Crenarchaeota
1
2
3
4
5
6
7
8
9
10
11
12
T
13
14
15
16
17
18
19
1 Thermatogales; 2 flavobacteria; 3 cyanobacteria; 4 purple bacteria; 5 Gram-positive bacteria;
6 green nonsulfur bacteria; 7 *Pyrodictium* sp.; 8 *Thermoproteus sp.*; 9 Thermococales; 10 Methanococcales;
11 Methanobacteriales; 12 Methanomicrobiales; 13 extreme halophiles; 14 animals; 15 ciliates;
16 green plants; 17 fungi; 18 flagelattes; 19 microsporidia
1 µm

Fig. 1. A. Universal phylogenetic tree in rooted form, as proposed by Woese *et al.*, (1990). The position of the order *Thermoplasmales* is marked with "T".
B. Electron microphotography of dividing *P. torridus* cells. The image was provided by Dr. O. Fütterer and Dr. M. Hoppert.

Since then, their phylogenetic position as a distinct group has also been confirmed by finding major differences in their plasma membrane lipid composition, cell wall, informational processes or metabolism (for a review see Kelly *et al.*, 1994; Reeve, 1999; Boucher *et al.*, 2003). The relatedness of archaea to the eukaryotes is thought to make them a suitable model for studying basic biological processes relevant to higher organisms. Also, the archaeal species that live in extreme environments give an opportunity to investigate the mechanisms of adaptation to these conditions and are considered to be a valuable source of biotechnologically important enzymes and macromolecules.

Picrophilus torridus is a moderately thermophilic, acidophilic archaeon that lives optimally at 60°C and pH 0.7. Strains of this species were first isolated from a dry solfataric field in northern Japan (Schleper *et al.*, 1995). In these geothermally heated habitats, the acidity is due to the sulphuric acid formed by the oxidation of volcanic sulphur to SO_3 which reacts with water to produce H_2SO_4. The acid concentration can be further increased by water evaporation. Two *Picrophilus* species have been described, *P. torridus* and *P. oshimae,* and they form a distinct family within the order of the *Thermoplasmales* of the euryarchaeal phylum (Fig. 1A).

In addition to being moderately thermophilic, the *Picrophilaceae* are the most acidophilic organisms known and are also able to grow at negative pH values. It has been reported for *P. torridus* that even adaptation to conditions such as in 1.2 M sulfuric acid is possible (Schleper *et al.*, 1995), and in the course of the current work this was repeatedly confirmed. The most acidophilic prokaryote identified previously was *Thermoplasma acidophilum* which grows optimally between pH 1.8 and 2 (Darland *et al.*, 1970). Interestingly, four eukaryotic organisms have been found to be able to survive at pH values around zero: a coccoid rhodophyte and three fungal species (Brock, 1978). Another archaeal group of thermoacidophilic organisms can be found in the crenarchaeal lineage, e.g. the *Sulfolobales*. The genus *Sulfolobus* comprises several species, most of which are able to gain their energy by oxidising sulphide to molecular sulphur. In contrast to the *Thermoplasmales*, they are very well studied due to their capability of gene transfer which has led to the development of several genetic systems.

P. torridus cells are irregular cocci of approximately 1 to 1.5 µm in diameter, enveloped in a 20 nm thick S-layer with a tetragonal symmetry and an additional brush-like structure on top of the S-layer, presumably consisting of long polysaccharide chains (Schleper *et al.*, 1995; Fig 1B). The organism grows heterotrophically and aerobically on 0.2 % yeast extract, and the addition of 1 % glucose or 0.2 % starch to the growth medium has been reported to lead to higher cell densities (Serour *et al.*, 2002). The *Picrophilus* species are unable to grow by fermentation which is in line with the observation that no fermentative acidophiles have been described so far (Johnson, 1998). The temperature and pH intervals permitting growth of *P. torridus* are 45-65°C and pH 0-3.5 respectively. The presence of a 8.8 kb plasmid has been shown in *P. oshimae*, but not in *P. torridus* (Schleper *et al.*, 1995). An unusual trait of *Picrophilus* sp. is a very low intracellular pH of 4.6 in contrast to other thermoacidophilic organisms which maintain internal pH values close to neutral (van de Vossenberg *et al.*, 1998). The high specialization of *Picrophilus* strains for growth in extremely acidic habitats is evident from their inability to grow at pH values above 4.0.

Considering the extraordinary growth conditions of *Picrophilus* sp., several characteristic adaptational features can be expected: i) As the cytoplasmic membrane is the only physical barrier against the acidic environment, it has to be able to withstand a steep pH gradient of 4-5 pH units. It has been shown in an *in vitro* system that liposomes derived from *P. oshimae* lipids have extremely low proton permeability at acidic pH and high temperature (van de Vossenberg *et al.*, 1998). ii) Specific modifications of the metabolism reflecting the low intracellular pH and the large pH gradient across the membrane are expected, the most relevant ones being the energy generation processes. iii) The structure and function of the organism's macromolecules should reflect their inherent stability at these conditions. It should be particularly interesting to study the extracellular proteins, i.e. S-layer protein, secreted enzymes etc. iv) Maintaining the pH homeostasis over a broad extracellular pH range requires specific regulation mechanisms at different levels. For example, well studied parasitic microorganisms which, in their life cycle, have to cope with acid environment like *E. coli* and *Helicobacter pylori* have been shown to possess pH dependent regulation at the gene expression level (Bearson *et al.*, 1997; Foster, 1999; Tucker *et al.*, 2002).

The whole genome sequence of an organism gives the unique opportunity to obtain information about its total set of genes and serves as a basis for a systematic study of the organism's biology. Since 1995 when the *Haemophilus influenzae* whole genome sequence was released (Fleischmann *et al.*, 1995), 190 genomes have been published, 18 of which belong to archaeal organisms (GOLD database as of April, 2004). Of these, although being among the first, only 4 are of thermoacidophiles, i.e. *T. acidophilum* (Ruepp *et al.*, 2000), *T. volcanium* (Kawashima *et al.*, 2000), *S. solfataricus* (She *et al.*, 2001), and *S. tokodaii* (Kawarabayasi *et al.*, 2001). The analysis of these genomes has led to major advances in the comprehension of the genetics and physiology of extremophilic organisms and has enhanced the understanding of the molecular basis of adaptation to extreme environments (for a review see Ciaramella *et* *al.*, 2002). The information obtained from the complete genome sequence of an organism is of particular significance when the organism is difficult to be cultured and/or is not amenable to genetic manipulation. With a few exceptions, among them *S. solfataricus* (Jonuscheit *et al.*, 2003), *Pyrococcus* sp. (Lucas *et al.*, 2002) and *Haloferax* sp.(Bitan-Banin *et al.*, 2003), this is most often the case in Archaea. In addition, the genome data has proven to be a valuable source of robust proteins and enzymes useful in biotechnological applications.

The major intention of the current work was to determine the complete genome sequence of *P. torridus* followed by analysis of the identified genes and the possible metabolic pathways which they could encode. The information obtained from the genome was further interpreted with respect to the unique characteristics of the organism. Also, it was anticipated that the genome sequence of *P. torridus* would allow a more complex investigation of the evolutionary relations among organisms that share similar extreme growth conditions under the special consideration of lateral gene transfer.

Another goal of the current work was the cloning and heterologous expression of selected *P. torridus* ORFs and the subsequent biochemical characterisation of the recombinantly produced proteins. A common obstacle in the analysis of archaeal proteins is that there are no suitable expression systems available. Problems in obtaining such proteins in a heterologous host can arise from the codon usage of the corresponding genes, the nascent protein folding system or the specific properties of the polypeptides, e.g. their inability to take a native conformation under “normal” conditions (Hartl *et al.*, 2002). Therefore, different approaches of obtaining recombinant proteins were tested in the course of the current work. Proteins chosen for investigation were biotechnologically important enzymes with enhanced stability at elevated temperature and low pH, e.g. glucose dehydrogenase and glycoside-hydrolysing enzymes. With the exception of a native glucoamylase, described by Serour *et al.* (2002), no other enzymes of *P. torridus* have been studied previously.

Glucose dehydrogenase: pyridine dependent glucose dehydrogenases oxidize glucose to gluconate via gluconolactone with the concomitant reduction of the cofactor $NADP^+$ or NAD^+. Enzymes with such activity have been isolated from higher eukaryotes and sporulating bacteria (Thompson *et al.*, 1970; Fujita *et al.*, 1977). The presence of pyridine nucleotide-dependent glucose dehydrogenases in archaea has been associated with the catabolism of glucose via a modified, non-phosphorylated Entner-Doudoroff pathway (Budgen *et al.*, 1986).

There is considerable interest in glucose dehydrogenase enzymes that are stable to heat, pH, organic solvents or proteolysis. They are used for the quantitative determination of glucose in blood and other fluids by a single-step assay, and are thus applied in clinical tests and in the food industry (D'Auria *et al.*, 2000). This has led to numerous attempts to increase the stability of mesophilic enzymes by different methods: site-specific mutagenesis, evolutionary mutagenesis, gene shuffling (Makino *et al.*, 1989; Baik *et al.*, 2003). Also, *T. acidophilum* has served as a source for the isolation of a glucose dehydrogenase with increased half-life at high temperatures and in organic solvents (Smith *et al.*, 1989).

α-Glucosidase: α-glucosidases are enzymes that typically catalyse the hydrolysis of terminal, non-reducing 1,4 linked D-glucose residues. In contrast to glucoamylases (glucan 1,4-alpha-glucosidases) α-glucosidases hydrolyse oligosaccharides rapidly compared to polysaccharides, which are hydrolyzed relatively slowly, or not at all. Numerous α-glucosidases have been characterised from bacteria and eukaryotes, the majority of them from mesophilic organisms. A thermoactive α-glucosidase from *Thermotoga maritima* with unusual cofactor requirements has been described recently (Raasch *et al.*, 2000). The reported archaeal representatives are limited to the enzymes from *Sulfolobus solfataricus* (Rolfsmeier *et al.*, 1995), *Pyrococcus furiosus* (Constantino *et al.*, 1990) and *Thermococcus* sp. (Piller *et al.*, 1996) of which only the *S. solfataricus* α-glucosidase has been recombinantly expressed (Haseltine *et al.*, 1999). The biotechnological potential of α-glucosidases that are active at high temperature and acidity and have a long shelf life lies in the bioprocessing of plant starch.

A. MATERIALS AND METHODS

1. Bacterial strains and growth conditions

1.1. Strains and plasmids

The bacterial strains used in the current work are described in Table 1. In Table 2, the basic plasmid vectors and constructs are summarised.

Table 1

Strain	**Description**	**Reference**
<u>*Picrophilus*</u>		
P. torridus	type strain (DSM 9790)	Schleper, C. *et al.*, 1996
P. oshimae	type strain (DSM 9789)	Schleper, C. *et al.*, 1996
<u>*Sulfolobus solfataricu*</u>s		
S. solfataricus P1	wild type (DSM 1616)	Zillig, W. *et al.*, 1980
S. solfataricus PH1-16	Δ *lac*S, *ura*	Schleper, C., 1994
<u>*Escherichia coli*</u>		
XL1-Blue	*recA*$^-$, *thi*, *hsdR*1, *sup*E44, *rel*A1, *lacF'*, *proAB*, *lacI*^{q}, *lacZ*ΔM15, Tn10[Tet]	Bullock *et al.*, 1987
TOP10	F- *mcr*A Δ*(mrr-hsd*RMS-*mcr*BC*)* *f80lacZ*ΔM15 Δ*lac*X74 *deo*R *rec*A1 *ara*D139 Δ*(ara-leu)*7697 *gal*U *gal*K *rps*L *end*A1 *nup*G	Bachmann, 1990
Stbl4	*mcr*A .(*mcr*BC-*hsd*RMS-*mrr*) *rec*A1 *end*A1 *gyr*A96 *gal- thi*-1 *sup*E44 λ- *rel*A1 .(*lac*-*pro*AB)/F. *pro*AB+ *lac*IqZ.M15 *Tn*10 (TetR)	Invitrogen Carlsbad, CA, USA
BL21 (DE3)	*hsdF*, *gal(λcI*ts857 *ind1* Sam7 *nin5* *lacUV5*-T7 gene*1)*	Studier and Moffat, 1986
Rosetta (DE3)	F- *ompT* *hsdS*$_B$(r$_b^-$ m$_b^-$) *gal dcm* *lacY1* (DE3) pRARE (Cm^R)	Novagen, Darmstadt, Germany.

Table 1 cont.

Saccharomices cerevisiae INVSc1	MAT**a** *his3.1 leu2 trp1-289 ura3-52*/MATб *his3.1 leu2 trp1-289 ura3-52*	Invitrogen Carlsbad, CA, USA

Phenotype description: Km^R-kanamycin resistant, Amp^R – ampicilin resistant, Cm^R – chloramphenicol resistant.

Table 2 Plasmids used. Detailed maps of the plasmids are shown in Appendix C

Plasmid	**Description**	**Reference**
Lorist6	Cosmid vector for genomic library construction, Kan^R	Gibson *et al.*, 1987
pBluescript II KS$^+$	High copy number cloning vector , oriEc (*colE*1), *lacPOZ'*, Ap^R	Stratagene, La Jolla, CA, USA
pCR4-TOPO	High copy number cloning vector, oriEc (*colE*1), *lacPOZα-ccdB*, Km^R, Ap^R	Invitrogen Carlsbad, CA, USA
pBAD/*Myc*-His	Expression vector, *araC*, P_{BAD}, *rrnB*, *oriEc* (*colE1*) Amp^R	Invitrogen Carlsbad, CA, USA
pET24 c/d	Expression vector, P_{T7}, *lacI*, pBR322 *ori*, Kan^R	Novagen Darmstadt, Germany.
pET 43.1	Expression vector, enables expression of proteins fused with the Nus Taq protein, P_{T7}, *lacI*, pBR322 *ori*, Kan^R, Nus taq	Novagen Darmstadt, Germany.
pMAL c2x	Expression vector, enables expression of proteins fused with the MBP from *E. coli*, P_{TAC}, *malE*, *lacZα*, *rrnB*, *lacI*^{q}, Amp^R	New England Biolabs, Beverly, MA, USA
pYes2 NT-A	Expression vector, allows expression of recombinant proteins in *Saccharomyces cerevisiae*, P_{GAL1}, *CYC1*, *URA3*, pUC *ori*, Amp^R	Invitrogen Carlsbad, CA, USA
pMJ03	*E. coli/S. solfataricus* shuttle viral vector, allows expression of heterologous proteins in *S. solfataricus*	Jonuscheit *et al.*, 2003

Table 2 cont.

pBAD-GDH	pBAD/*Myc*-His expression vector carrying the *P. torridus gdh*A gene under the control of the P_{BAD} promoter.	this work
p24-GDH	pET24d expression vector carrying the *P. torridus gdh*A gene under the control of the P_{T7} promoter.	this work
pBAD-1383	pBAD/*Myc*-His expression vector carrying the *P. torridus* ORF RPTO 01383 under the control of the P_{BAD} promoter	this work
p24-1383	pET24d expression vector carrying the *P. torridus* ORF RPTO 01383 under the control of the P_{T7} promoter	this work
pMAL-1383	pMAL c2x expression vector carrying the *P. torridus* ORF RPTO 01383 under the control of the P_{TAC} promoter	this work
p24-615	pET24d expression vector carrying the *P. torridus* ORF RPTO 00615 under the control of the P_{T7} promoter	this work
pBAD-810	pBAD/*Myc*-His expression vector carrying the *P. torridus* ORF RPTO 00810 under the control of the P_{BAD} promoter	this work
p24-810	pET24d expression vector carrying the *P. torridus* ORF RPTO 00810 under the control of the P_{T7} promoter	this work
pMAL-810	pMAL c2x expression vector carrying the *P. torridus* ORF RPTO 00810 under the control of the P_{TAC} promoter	this work
pNus-810	pET43.1 expression vector carrying the *P. torridus* ORF RPTO 00810 fused with the *E. coli* Nus protein under the control of the P_{T7} promoter	this work
pMJ-810	pMJ03 shuttle vector carrying the *P. torridus* ORF RPTO 00810 under the control of the *S. solfataricus tf55*α promoter	this work
pYes-810	pYes2 NT-A expression vector carrying the *P. torridus* ORF RPTO 00810 under the control of the *S. cerevisiae GAL1* promoter.	this work
p24-596	pET24d expression vector carrying the *P. torridus* ORF RPTO 00596 under the control of the P_{T7} promoter	this work
p24-985	pET24d expression vector carrying the *P. torridus malP* gene under the control of the P_{T7} promoter	this work

Phenotype description: P_{BAD} - *ara*BAD promoter from *E.coli*, P_{tac} – fused *trp* / *lac*UV5 promoter (Amann *et al.*, 1983), *rrnB*- replication termination region from *rrnB* operon in *E. coli* (Brosius *et al.*, 1981), *gdh*A – *P. torridus* glucose dehydrogenase gene, *mal*P – *P. torridus* alpha–glucosidase gene.

1.2. Growth media

Liquid media were prepared in bidistilled water and autoclaved at 120 °C for 20 min. Solid media were prepared with the addition of 14 g/l bacteriological agar (Oxoid, Wesel, Germany) before autoclaving. Substrates that are sensitive to autoclaving such as antibiotics or sugars were sterilized by filtration (0.2 µm, Minisart, Sartorius, Goettingen, Germany) and added to the media after autoclaving at a medium temperature lower than 60 °C.

E. coli strains

- **LB medium** (Sambrook *et al.*, 1989)

tryptone	10 g
yeast extract	5 g
NaCl	5 g
dd H_2O	up to 1000 ml

When required, antibiotics, IPTG, X-gal or other supplements were added after autoclaving at concentrations described in Table 3.

Table 3. Media additives

Additive	Abbr.	Stock solutions	Final concentration in the media
Ampicillin	Amp	50 mg/ml in water	50 µg/ml
Kanamycin	Km	10 mg/ml in water	20 µg/ml
Chloramphenicol	Cm	25 mg/ml in ethanol	12 µg/ml
Isopropyl-ß-d-thiogalactopyranosid	IPTG	100 mM in water	0.2 mM
5-Brom-4-chlor-3-indolyl-ß-D-galactopyranosid	X-gal	20mg/ml in DMF	40 µg/ml
L-arabinose	L-ara	20% (w/v) in water	0.2 % (w/v)

The stock solutions were sterilized by filtration, aliquoted in 1.5 ml volumes and stored at -20°C.

S. cerevisiae

- **SC minimal medium**

SC is a synthetic minimal defined medium for yeast. In the current work the strain INVSc1 was used, which is a histidine/tryptophan autotroph, and these aminoacids were supplemented in the medium.

Yeast nitrogen base	6.7 g
Tryptophan	0.1 g
Histidine	0.05 g
D-Glucose	20 g
Agar (for preparing plates)	20 g

All the components (except agar when preparing plates) were prepared as stocks and sterilised separately.

Picrophilus strains

- **DSMZ 723 medium**

$(NH_4)_2SO_4$	1.30 g
KH_2PO_4	0.28 g
$MgSO_4$ x 7 H_2O	0.25 g
$CaCl_2$ x 2 H_2O	0.07 g
$FeCl_3$ x 6 H_2O	0.02 g
$MnCl_2$ x 4 H_2O	1.80 mg
$Na_2B_4O_7$ x 10 H_2O	4.50 mg
$ZnSO_4$ x 7 H_2O	0.22 mg
$CuCl_2$ x 2 H_2O	0.05 mg
Na_2MoO_4 x 2 H_2O	0.03 mg
$VOSO_4$ x 2 H_2O	0.03 mg
$CoSO_4$	0.01 mg
Yeast extract (Difco)	2.00 g
0.5 M H_2SO_4	300 ml
Distilled water	up to 1000.00 ml

Yeast extract was autoclaved separately and added to the medium last. The use of diluted sulphuric acid resulted in a pH of the medium of about 1, which was further adjusted with concentrated sulphuric acid to pH 0.5 at 60°C.

S. solfataricus strains

- **DSMZ 182 medium**

Yeast extract (Difco)	1.00 g
Casamino acids (Difco)	1.00 g
KH_2PO_4	3.10 g
$(NH_4)_2 SO_4$	2.50 g
$MgSO_4$ x 7 H_2O	0.20 g
$CaCl_2$ x 2 H_2O	0.25 g
$MnCl_2$ x 4 H_2O	1.80 mg
$Na_2B_4O_7$ x 10 H_2O	4.50 mg
$ZnSO_4$ x 7 H_2O	0.22 mg
$CuCl_2$ x 2 H_2O	0.05 mg
Na_2MoO_4 x 2 H_2O	0.03 mg
$VOSO_4$ x 2 H_2O	0.03 mg
$CoSO_4$ x 7 H_2O	0.01 mg
Distilled water	up to1000.00 ml

The pH was adjusted to 3.5-4 with 10 N sulphuric acid.

1.3. Growth conditions

E. coli and *S. cerevisiae* strains were cultured both in liquid and on solid media. Liquid cultures were grown in Erlenmeyer flasks with medium volumes representing maximally 1/10 of the flask volume. An exception was made for 5 ml liquid cultures for which standard test tubes were used. For optimal aeration the Erlenmeyer flask cultures were incubated on a flat-deck rotary shaker while the test tubes were agitated on racks with fixed 40^o angle to the shaking surface at 150 rpm. Cultivation on solid media was performed on 92 mm disposable plastic plates (Sarstedt, Nümbrecht, Germany) prepared as follows: 1.4 % (w/v) agar was added to the liquid media before autoclaving. Autoclaved agar-containing media were left to cool down to 60 ^{o}C and, after supplementation of the additives, were poured into sterile plates. Growth temperatures were 37°C or 30°C for *E. coli* and 30°C for *S. cerevisiae*.

When *P. torridus* and *S. solfataricus* were grown, precautions had to be taken with respect to the elevated growth temperatures. *P. torridus* was incubated in a Innova 4400 (New Brunswick Scientific, Madison, NJ, USA) incubator/shaker at 60°C, and for *S. solfataricus* a rotary water bath shaker at 75°C was used, filled with non-evaporating liquid - Rotitherm K+H, (Carl Roth GmbH, Karlsruhe, Germany). The Erlenmeyer flasks used for culturing *S. solfataricus* were with elongated bottlenecks (25 cm) in order to prevent excessive loss of liquid via evaporation.

1.4. Monitoring of growth, culture harvesting and cell fractionation

The growth of unicellular organisms could be rapidly determined using the linear dependence of the turbidity of the cell suspension to the cell number. The cultures' turbidity was quantified by photometric measurement of the optical density at 600 nm (OD_{600}) against the pure growth medium as a blank (Pharmacia, Uppsala, Schweden). When necessary, the cell suspensions were diluted in order to keep the OD_{600} measurements in the range of 0.1-0.4, where there is a linear dependence between the optical density and the number of cells.

The cells of a growing culture, after cooling on ice, were collected by centrifugation (6000 g, 4 ^{o}C, 15 min), and washed with a suitable buffer. All subsequent fractionation steps were carried out on ice.

Sonication was routinely used as a method for cell disruption, using a stationary sonicator (Dr. Hielscher, Stahnsdorf, Germany); when larger volumes were processed, the cells were opened by passing them twice through a French Press Cell. *S. cerevisiae* cells were opened by vortexing the cell suspension with a small amount of glass beads for 30 sec, cooling on ice for 1 min and repeating this treatment for 5 times.

1.5. Storage of strains and control of purity

Permanently used *E. coli* strains were maintained on LB selective agar plates, that could be stored at 4 °C up to 2 months. For long-term storage bacterial strains were stocked as glycerol cultures at –70 °C. For this purpose, 0.7 ml of fresh culture, grown overnight in complex media in the presence of selective pressure as required, was mixed with an equal volume of 50 % (w/v) glycerol sterilized by autoclaving. Stock cultures, prepared in this way are stable over years with moderate loss of viability. In order to control the purity, the strains were propagated on both selective and non-selective agar plates and checked for the uniformity of the colonies. The plasmid - containing strains were additionally checked by isolation and analytical restriction of the plasmids. When preparing glycerol stock cultures from *P. torridus*, the pH of the culture was adjusted to 3.5 with sterile filtered 0.5 N NaOH before addition of glycerol.

2. DNA manipulations

2.1. General techniques

All tools, vessels and solutions for work with DNA were autoclaved (20 min, 120°C) for inactivation of DNA-degrading enzymes. Tools that are not autoclavable were first rinsed with 70 % (w/v) ethanol and subsequently with sterile dd H_2O. Non-autoclavable or heat-unstable substances (e.g. lysozyme, proteinase K) were dissolved in sterile buffers or water.

2.1.1. DNA purification

<u>Phenol- chloroform extraction</u>

- Phenol-chloroform solution**:**

phenol(pH 8.0)/chloroform/isoamyl alcohol	25:24:1 (v/v/v)

- Chloroform-isoamyl alcohol solution**:**

chloroform/isoamyl alcohol	24:1 (v/v)

The phenol-chloroform extraction removes protein contaminations from DNA-containing samples, using the different behavior of proteins in comparison to DNA during extraction with organic solvents. While the proteins are denatured by the organic solvent, the DNA remains in soluble form and can be recovered from the aqueous phase.

The DNA-containing solutions (bacterial extracts, restriction reaction mixtures etc.) were mixed with an equal volume of phenol-chloroform and the two phases were mixed by shaking. The phases were then separated by centrifugation (16,000 x *g*, 5 min, RT). The upper aqueous phase was carefully pipetted out in a new tube, trying to avoid any contact with the protein interlayer. For viscous chromosomal DNA-containing mixtures, pipette tips with cut ends were used. The procedure was repeated several times.

The chloroform-isoamyl alcohol extraction was performed as a stand-alone DNA purification step, or as a last step of the phenol-chloroform extraction procedure. This procedure removes trace amounts of phenol from the DNA solution, which can cause problems in further DNA manipulations. The DNA solution was mixed with an equal volume of chloroform-isoamyl alcohol and mixed vigorously. The mixture was centrifuged (16,000 x *g*, 5 min, RT) and the upper aqueous phase was carefully transferred into a new tube. The purified DNA in this phase was used for further manipulations either directly or after alcohol precipitation.

Isopropanol /ethanol precipitation

The precipitation of DNA with ethanol or isopropanol was used for the purification and concentration of DNA samples. The precipitation with isopropanol has the advantage of reduced volume, but in contrast to ethanol precipitation results in a transparent DNA pellet which could sometimes be problematic in further manipulations.

The DNA-containing samples were mixed with 0.7 volumes isopropanol or 2.5 volumes absolute ethanol. The mixtures were incubated 15 min on ice (alternatively, 5 min at –70 ^{o}C) and precipitated DNA was pelleted by centrifugation (16,000 x *g*, 15 min, 4 ^{o}C). The pellet was washed with 70 % (v/v) ethanol and centrifuged again after 15 min incubation on ice (16,000 x *g*, 15 min, 4 ^{o}C). The remaining ethanol was carefully pipetted out and the pellet was dried at 37 ^{o}C for 5-10 min to allow the evaporation of the remaining ethanol. Longer incubations result in overdrying of the

DNA, that leads to decreased solubility. Dried DNA was finally dissolved in TE buffer or sterile water.

Isolation of DNA from agarose gels

DNA fragments obtained from restriction digestion or PCR amplification were separated on an agarose gel and isolated from it. The procedure allows the isolation of DNA fragments with defined molecular size from a mixture of linear DNA molecules. After staining a gel containing the DNA fragments of interest (see 2.1.2.), parts of the gel containing these fragments were cut out and used for DNA isolation with the QIAquick Gel Extraction Kit (QIAGEN, Hilden, Germany). The extraction procedure was made according the manufacturer’s instructions and the DNA was eluted from the column using 50 µl sterile dd H_2O. Optionally, the DNA solution was concentrated after the extraction procedure using a vacuum evaporation centrifuge SpeedVac Plus (Sevant). The purity and the concentration of the isolated fragments were checked on an analytical agarose gel.

2.1.2. DNA analysis using agarose gel electrophoresis

- Tris-acetate-EDTA (TAE) buffer

Tris-acetate	40 mM (pH 8.1)
EDTA	2 mM

A 50x solution was made, which was diluted before use

- Loading buffer

Glycerol	30 % (v/v)
EDTA	50 mM
Bromphenol blue	0.25 % (w/v)
Xylene cyanol	0.25 % (w/v)

The DNA electrophoresis was performed in a horizontal mini gel apparatus (Hoefer HE33, Pharmacia) with a gel size of 10 x 6.6 x 0.8 cm. Corresponding combs enabling the creation of 10 or 16 slots per gel were used. The concentration of the agarose gel varied between 0.5 and 1.2 % (w/v) agarose, depending on the size of the DNA fragments that should be separated. The electrophoreses were run in 1x TAE buffer at a constant voltage of 100 V (BioRad Power Pac 300 power supply) for 30-60 min. For visualization of the DNA the gels were incubated in an ethidium bromide solution (1.5 µg/ml water) for 5-10 min, followed by washing with ddH_2O. The DNA was visualized under UV light and the DNA fragment profiles were documented using a GelDoc system (BioRad, München, Germany). The size of the DNA fragments was evaluated using 1 kb DNA and 100 bp DNA ladder (MBI Fermentas) size-marker standards run on the same gel.

2.1.3. DNA quantification

Spectrophotometric quantification

The concentration of DNA was measured using the property of DNA to absorb UV light with a maximum at 260 nm. The concentration of double strand DNA is proportional to the OD_{260} in the range of 0.1-0.8 (Cryer *et al.*, 1975), where OD_{260} of 1 corresponds to 50 µg/ml DNA concentration. Additional information about the purity of the DNA could be gained by the ratio OD_{260}/OD_{280}. Values higher than 1.5 indicate a high degree of purity of the DNA sample.

Estimation from agarose gels

Alternatively, the amount of DNA was estimated by comparison of the bands' intensity after agarose gel separation with marker bands of corresponding size and defined concentration. This quantity estimation method was also used for the direct comparison of the DNA amounts in different samples.

2.2. Isolation of DNA

2.2.1. Plasmid isolation from *E. coli*

For plasmid isolation *E. coli* cells were grown in 5 ml LB media with selective pressure. The cells from 4 ml culture were harvested in 2 ml reaction tubes by subsequent roumds of centrifugation (16,000 x *g*, 5 min each, RT). The pellets were resuspended in 250 µl resuspension buffer (P1). Ineffective pellet resuspension can decrease the final yield by preventing efficient cell lysis. The cell suspensions were mixed with 250 µl lysis buffer (P2). The clearing of the mixture is an indication for efficient cell lysis. 350 µl of the neutralisation solution (P3) were rapidly added to the samples and mixed gently by inverting the tubes. The precipitate was removed by centrifugation (16,000 x *g*, 10 min, RT). The supernatant containing the plasmid DNA was further processed via isopropanol/ethanol precipitation.

For obtaining high quality plasmid DNA for sequencing, the plasmid DNA was purified using the QIAprep Plasmid Purification Kit (QIAGEN). The purification procedure was carried out according to the manufacturer's instructions. The pellets obtained by alcohol precipitation were dissolved in 50 µl of sterile dd H_2O by incubation at 37 ^{o}C for 10-30 min. The isolated plasmids were analyzed by restriction digestion and agarose gel electrophoresis.

2.2.2 Isolation of genomic DNA from *P. torridus*

- TE-Sucrose (*sterile filtered*)**:**

Sucrose	20 % (w/v)
in TE-buffer	

- Lysozym-RNase (*prepared daily*)**:**

Lysozym	100 mg/ml
RNase sol.	1 % (v/v)
in TE-buffer	

- N-Lauryl-Sarcosine-Proteinase K(*prepared daily*)**:**

N-Lauryl-Sarcosine	5 % (w/v)
Proteinase K	~ 1 mg/ml
in dd H_2O	

The method was used for the isolation of genomic DNA from *Picrophilus sp.* Five ml from 40 ml cultures grown in complex media were harvested by centrifugation (16,000 x *g*, 10 min, RT). The pellets were dissolved in 250 µl TE-Sucrose buffer and 250 µl of the Lysozyme-RNase solution. After 1 hour of incubation at 37 ^{o}C, 250 µl of

N-Lauryl-Sarcosine-Proteinase K solution were added. The mixtures were incubated for at least 1 h at 37 °C. High viscosity and transparency of the solution after this step indicates efficient cell lysis. The solutions were passed through a 1 ml pipette tip several times and subjected to phenol-chloroform extraction followed by isopropanol precipitation. The dried DNA pellets obtained were dissolved in 50-100 µl water or TE-buffer by overnight incubation at 4 °C. The purity and quality of the chromosomal DNA were analyzed by agarose gel electrophoresis with and without subjecting it to restriction digestion.

2.3. Enzymatic modification of DNA

2.3.1. Restriction

Analytical scale digestion with restriction enzymes was performed for the characterization of different DNA constructs.

- Analytical digestion reaction**:**

DNA solution	up to 1 µg
10x Restriction buffer	2 µl
Restriction enzyme	2-5 U
dd H_2O	up to 10 µl

The digestions were performed for 1 h at the temperature optimal for the restriction enzyme’s activity. The results of the digestion reaction was directly analyzed by agarose gel electrophoresis.

If one or more of the fragments resulting from the restriction reaction were to be used in further cloning procedures, the reaction volumes were scaled up as follows:

- Preparative digestion reaction**:**

DNA solution	max. 10 µg
10x Restriction buffer	5 µl
Restriction enzyme	10-25 U
dd H_2O	up to 50 µl

and the digestions were performed for 2 h at the temperature optimum of the enzyme. If the digestion was not completed in this period, additional 10 U of the enzyme were added and the reaction was continued for another 2 hours. When digesting with two ore more enzymes, the universal buffer systems OPA (Amersham-Pharmacia) or TangoY (MBI Fermentas) were used. If the enzymes used did not have sufficient activity in a common buffer system or under common conditions, the digestions were made in separate steps with DNA fragment purification and buffer exchange between the steps.

2.3.2. Dephosphorilation of linearised DNA

In order to avoid re-ligation of empty plasmid vectors during the ligation reaction, the 5`-phosphate residues at the end of linearized vectors were removed by alkaline phosphatase treatment. After preparative vector DNA digestion, the reaction volume was adjusted to 120 µl, 15 µl 10x dephosphorylation buffer (Boehringer Mannheim, Germany) and 15 U shrimp alkaline phosphatase (Boehringer Mannheim) were added and the sample was incubated at 37 °C for additional 2 h. The phosphatase was inactivated by incubation at 65 ^{o}C for 20 min. The dephosphorylated DNA was purified by chloroform-isoamyl alcohol extraction followed by ethanol precipitation and was finally dissolved in 10 µl dd H_2O.

2.3.3. Ligation

For the ligation of DNA fragments into plasmid vectors the bacteriophage T4 ligase was used. The ligation represents a competitive reaction between intramolecular religation of both ends of one DNA molecule and ligation of the ends of two different molecules. This indicates the importance of the molar ratio of the different DNA molecules present in the ligation mixture. When aiming to introduce a foreign DNA fragment (insert) into the linearized plasmid vector, the molar concentration chosen for the insert DNA exceeded the vector DNA concentration at least 3-fold. When dephosphorylated vector was used or the size of the insert was significantly smaller than the size of the vector, an insert : vector molar ratio of 1:1 and 1:3, respectively, was used.

The ligation mixture was prepared on ice as follows:

- Ligation reaction**:**

DNA of the vector	x µl (0.1-0.2 µg)
DNA of the fragment	x µl (0.3-1 µg)
10x T4-DNA ligase buffer	1 µl
T4-DNA Ligase	1 µl (1 U)
dd H_2O	up to 20 µl

The ligation was carried out at 16 °C overnight.

2.4. *In vitro* DNA amplification. Polymerase chain reaction (PCR)

PCR reactions were used for the *in vitro* amplification of DNA fragments for cloning and analytical purposes. A PCR reaction was done using the thermostable DNA polymerase enzymes *Taq* (own preparation, data not shown) and *Pfu* (Promega, Mannheim, Germany). All the primers used in PCR amplifications for cloning purposes are listed in Table 4. The primers were designed to have between 18 and 25 bp homology with the target sequence and GC contents between 40 and 60 % if possible.

Table 4. PCR primers used for cloning. The primers used for closing genome gaps and analytical purposes are not listed.

Primer name	**Sequence**	**T_{ann}.**
517Afor	CATATGCATATAAGATTTATCAATGGTTTTATG	59
517rev	ATTCAGGCTCCTCCATGCCAATC	66
539for_nde	CACTGGAGGTTTACAGATCCATATGTCGCATGG	71
539rev	TGCCCAACAGGAAAATGTGATC	63
1070for_nde	GTGCATATGTTACCCAAGAACTTTTTAC	63
1070rev	ACGTTCTCTGAAGTAGCCTTGCC	66
seq1070	TATCCTGCTCCATTCAACTC	58
421for	GGCGTTCATAACCCTTGTTACCTCTTCA	68
421rev	CGTCATGCCATCAACGTCCTTGTAGAAT	68

Table 4 cont.

1383for_nde	CTCATATGGAGACAATAAAAAGCGTAGA	63
1383rev	GAGAATGGGAACCTAAAGGATGAG	62
S1070F-nco	TGGCCATGGGCTTACCCAAGAACTTTTTACTTG	72
S1070R-eag	GCCGGCCGCTCATATGGCCAATTATAAAG	71
TF55-1070.R	GTTCTTGGGTAACATGACTGGAGCTGCCATACC	73
pyrEF.F	CTGGATCCCAGCAGACGTATAAAAGCC	68
985reg_for	TGCGGAATACCATTCGGCAGCAT	69
985reg_rev	TCAATACGGCCGCACCAACAAGT	69

2.4.1. Analytical PCR

For the confirmation of a bacterial strain genotype or for checking different DNA constructs, an analytical PCR reaction was used. The PCR was performed using *Taq* polymerase (*Termus aquaticus*) heterologously expressed in *E. coli* (own preparation). The *Taq* enzyme has the disadvantage of higher error frequencies (8.0 10^{-6} errors per base per duplication) but has high processivity, which makes it suitable for analytical PCR purposes.

The PCR reaction was made in 0.2 ml plastic tubes with reaction volumes between 20 and 100 µl. For several or many parallel PCR reactions, a reaction master mix was prepared as follows :

- *Taq* PCR reaction:

10x Taq buffer	10 µl
dNTP mix (10 mM each)	2 µl
primer A (100 pmol/µl)	1 µl
primer B (100 pmol/µl)	1 µl
Taq polymerase	1 µl
dd H_2O	84 µl
Template DNA (max 0.5 µg/µl)	1 µl

- PCR conditions**:**

Initial denaturation	95 ^{o}C	5 min.
Tree-step cycle:		
Denaturation	95 ^{o}C	1 min.
Annealing	(T$_{ann.}$-5) ^{o}C	1 min.
Elongation	72 ^{o}C	1min. / kb
Number of cycles	25-35	
Final elongation	72 ^{o}C	10 min.
Store	4 ^{o}C	

The obtained PCR products were analyzed by agarose gel electrophoresis.

2.4.2. Preparative PCR

When the PCR products were used in further cloning reactions, a preparative PCR amplification was performed. In the preparative PCR, *Taq* enzyme was substituted by *Pfu* enzyme (Promega) – a high fidelity DNA polymerase from the thermophilic archaeon *Pyrococcus furiosus*. *Pfu* polymerase possesses a proofreading 3’-exonuclease activity that significantly decreases its error rate (1.3 x 10 $^{-6}$ error per base per duplication) in comparison to the error rate of the *Taq* enzyme. The use of this enzyme significantly decreases the probability for mutations introduced in the amplified fragment during the PCR reaction. The procedure used for *Pfu*-PCR reaction was:

- *Pfu* PCR reaction:

10x Pfu buffer (Promega)	10 µl
dNTP mix (10 mM each)	2 µl
primer A (100 pmol/µl)	1 µl
primer B (100 pmol/µl)	1 µl
dd H_2O	84 µl
Template DNA (max 0.5 µg/µl)	1 µl
Hot start: *Pfu* polymerase (Promega)	1 µl (3 U)

One of the critical factors in the *Pfu* polymerase reaction is the presence of proofreading (3’) as well as 5’ exonuclease activity, that could cause significant primer degradation mostly prior to the initial primer-template annealing step. To minimize this effect a hot-start PCR procedure was used. The simplified hot-start procedure was performed as follows: the reaction mixture was prepared without the addition of *Pfu* polymerase; then the PCR reaction was started and once the initial denaturation temperature was reached the *Pfu* polymerase was added to the reaction by pausing the thermocycler machine.

The PCR products were further purified and cloned as described in chapter B.2.4.4.

2.4.3. Colony PCR

Taq-based PCR was used for direct analysis of bacterial clones. In this case instead of a purified DNA, bacterial cells were directly used as a template for the PCR reaction. The method relies on the fact that initial incubation at 95 ^{o}C leads to partial cell lysis, making intracellular plasmid or genomic DNA available for PCR amplification. *Taq* enzyme was preferred for these amplifications. The reaction conditions and the mixture composition did not differ from the conditions of standard analytical PCR except for a prolonged initial denaturation step. The reaction master mix was prepared without the addition of template DNA and 20µl samples of the mixture were aliquoted in 0.2 ml PCR tubes. The bacterial clones were picked with the tip of 20 µl pipette tips and transferred to the reaction tubes by washing the tip several times in the PCR mixture. The PCR was performed as described in B.2.4.1 and the products were analyzed on agarose gels.

2.4.4. PCR purification and cloning

The specific PCR products obtained by preparative PCR were directly purified from the reaction mixture using the QIAquick PCR Purification Kit (QIAGEN). When high unspecific background was present, the products of the reaction were separated on an agarose gel and the products of interest were purified using QIAquick Gel Extraction

Kit (QIAGEN). The purified PCR products were cloned into plasmid cloning vectors using one of the following methods:

- Blunt-end cloning

In this case the blunt end PCR products produced by *Pfu* DNA polymerase were directly cloned into pBluescript KSII restricted with *Eco*RV. For this purpose, the purified PCR products were concentrated to 1/5 of the initial volume by vacuum evaporation using a SpeedVac Plus centrifuge. The concentrated PCR products were directly added to the ligation mixture, ensuring a high excess of the PCR products over the linearized vector. The ligation reaction was performed as described in chapter B.2.3.3. and insert-containing clones were selected using α-complementation (“blue-white”) screening (Sambrook *et al.*, 1989).

- Topoisomerase cloning (TOPO cloning)

TOPO cloning is based on the ability of Topoisomerase I from *Vaccinia* virus to create 3’ T -protruding ends by cleaving after the sequence CCCTT. The enzyme remains covalently bound to the 5’end of the cut DNA (a cloning vector possessing a TOPO site). A PCR product having 3’A-protruding ends leads to the liberation of the enzyme and covalent binding of the PCR product to the cloning vector.

TOPO cloning of PCR products was performed with a TOPO TA cloning kit (Invitrogen). For this purpose, the blunt ended *Pfu* PCR product was subjected to post-amplificational addition of 3’ A-overhangs, using *Taq* DNA polymerase as described in the manufacturer’s instructions. Four µl of the PCR mix after *Taq* incubation were subjected to the TOPO-cloning procedure and subsequently transformed into *E. coli* following the manufacturer’s instructions.

2.5. Transformation

2.5.1. Transformation of *E. coli*

The method used for the transformation of *E. coli* with plasmid DNA was based on the incubation of chilled cells and DNA in a solution containing Ca^{2+}, Rb^{+} and Mn^{2+}

ions, followed by a short heat shock treatment (Hanahan, 1985). The competent *E. coli* cells were prepared by incubation in Ca $^{2+}$, Rb^+ and Mn^{2+} solutions, aliquoted (100 µl) and stored at –70 ^oC. Before the DNA transformation, aliquots of the cells were thawed, the DNA was added and the transformation was induced by a short heat shock at 42°C for 60 sec. Then the cells were mixed with 700 µl of prewarmed (37°C) LB medium, incubated for 1 h at 37°C and plated on selective medium.

2.5.2. Transformation of *S. solfataricus*

In this work, *S. solfataricus* was transformed (transduced) with the shuttle viral-based vector pMJ03 and its derivatives (Jonuscheit *et al*., 2003). This vector system is based on the *Sulfolobus shibatae* SSV1 virus which, upon infection of *Sulfolobus* cells is stably integrated in the host chromosome (Schleper *et al*., 1992). The shuttle vector pMJ03 and the derivate, constructed in this work (pMJ-1070) were introduced into *S. solfataricus* cells by electroporation, as described by Schleper *et al*., 1992.

- the cells of a fresh 50 ml overnight culture were cooled on ice for 15 min, collected by centrifugation (6,000 x *g* at 4°C) and washed gradually in 50, 25 and 1 ml ice-cold 20 mM sucrose solution in order to remove the salts, present in the medium. The cell density was adjusted to 10^{10} cells/ml with 20 mM sucrose and were kept on ice until electroporation. Fifty µl competent cells were mixed with 1 µl dialysed DNA (maximally 300 ng) and subjected to electroporation using the following parameters:

Voltage	1.5	kV
Capacity	25	µF
Resistance	400	Ω

Under these conditions, the highest transformation efficiency is achieved when the resulting time constant is 9.1 msec (Schleper *et al.*, 1992). Immediately after the electroporation the cells were mixed with 1 ml growth medium (see chapter B.1.2.), incubated for 1h at 75°C and finally transferred to a preheated 50 ml culture.

2.5.3. Transformation of *S. cerevisae*

Transformation of *S. cerevisae* was accomplished as described by Elble, 1992:

Transformation buffer		
PEG 3350	40	% v/v
Lithium acetate	0.1	M
Tris pH 7.5	10	mM
EDTA	1	mM
DTT	0.1	mM

- 500 µl from a fresh overnight culture was pelleted by centrifugation, the supernatant removed and the cells mixed with 100 µg carrier salmon sperm DNA and 1 µg of the DNA to be transformed. 500 µl freshly prepared transformation buffer was added, the cell suspension was vortexed and left at RT overnight. After a 10 min heat shock (42°C) the cells were washed with water and plated on selective medium.

2.6. Methods used in genome sequencing, assembly and sequence analysis

The strategy used in this study for sequencing of the genome of *P. torridus* was whole genome shotgun sequencing (Venter *et al*., 1995). This approach has become routine in sequencing of small genomes and can be described with the following stages:

1) generation of a randomly represented genome small insert library, with insert sizes ranging from 2 to 3 kb.

2) sequencing of a substantial number of clones, necessary to generate a redundant coverage of the whole genome, using universal primers complementary to the ends of the cloning vector.

3) automated assembly of the generated sequences which results in the formation of “contigs” of overlapping sequence reads.

4) gap closure stage in which different strategies are applied to bridge the gaps between the contigs assembled in the previous stage.

2.6.1. Generation of a whole genome shotgun library

Shotgun library construction of the *P. torridus* genome and sequencing of the clones was accomplished by Integrated Genomics (Chicago, IL, USA).

2.6.2. Genome assembly and closing of gaps

The generated sequence data (trace files) was further processed using the Staden software package (Staden *et al.*, 2000). The package contains several programs for processing, assembly and editing of sequence data. An extensive documentation about the programs can be found at http://staden.sourceforge.net/documentation.html.

All sequenceces were assembled into contigs with the PHRAP assembly program (Ewing *et al.*, 1998) and edited with GAP4 of the Staden software package. Gap closure was accomplished by primer walking on plasmids originating from the library and by PCR reactions with genomic DNA as template. Gene and gene order comparisons with already sequenced genomes served as a verification for the assembly of the contigs. Additionaly, multiplex combinatorial PCR was implemented as an alternative in the closing of gaps (for a detailed description of the method see Tettelin *et al*., 1999).

2.6.3. Sequence analysis and annotation

Open reading frames (ORFs) likely to code for proteins were predicted by the YACOB software package (Tech *et al.*, 2003), based on the algorithms CRITICA (Badger *et al.*, 1999), ORPHEUS (Frishman *et al.*, 1998) and GLIMMER (Delcher *et al.*, 1999). Automatic and manual annotation was carried out with the ERGO annotation tool (Integrated Genomics), which was refined by searches against the Pfam, PROSITE, ProDom and COGs databases. Additionally, BLASTP (Altschul, 1990) searches in Swissprot, NR and TCDB databases were used in the annotation process.

The prediction of the origin of replication was based on purine and keto excess plots, the identification of repeats in intergenic regions with the program REPuter (Kurtz and Schleiermacher, 1999), and manual gene analysis.

For gene comparison, homology was specified as 30 % amino acid sequence identity. Ortho- and paralogous sequences were counted only once. The threshold for specifying genes into the categories archaeal, bacterial, eukaryotic or thermoacidophilic was set at an e-value of $1e^{-05}$ at the amino acid sequence level.

3. Protein manipulations and biochemical methods

3.1. Determination of protein concentration

The concentrations of proteins in the crude cell extracts or during a purification process were assayed using the Bradford method (1976, modified). 5-20 µl of appropriately diluted protein solution were added to 1 ml Bradford reagent (Biorad, Hercules, CA, USA) in 1 ml disposable plastic cuvettes (Sarstedt, Nümbrecht, Germany). After 5 min incubation at room temperature the absorption was measured at 595 nm with an Ultrospec 3000 spectrophotometer (Pharmacia) using pure Bradford reagent as a blank. The protein concentration was estimated based on the linear dependence between the OD_{595} and the protein concentration. A standard curve made up with 0 to 10 µg BSA was used as a reference.

3.2. Polyacrilamide gel electrophoresis (PAGE)

3.2.1. SDS-PAGE

SDS-polyacrylamide gel electrophoresis (SDS-PAGE) was used to analyse the protein composition of complex protein mixtures such as crude cell extracts or to analyse the behaviour of proteins during purification or refolding experiments. The technique is based on separation of proteins according to their molecular weights.

30 % (w/v) Acrylamide / Bis:	*30 % acrylamide / bisacrylamide* (*37.5/1*) *in water, premixed* (*Roth, Karlsruhe, Germany*)
0.5 M Tris-HCl (H 6.8):	*in water, autoclaved, stored at 4 °C*
1.5 M Tris-HCl (pH 8.8):	*in water, autoclaved, stored at 4 °C*

10 % (w/v) SDS: *in water*

2 % (w/v) bromophenol blue: *in water, stored at -20 ^{o}C*

sample buffer (4x):	0.5 M Tris-HCl (pH 6.8)	6.6 ml
	glycerol	7.5 ml
	10 % (w/v) SDS	12 ml
	2 % (w/v) bromophenol blue	0.5 ml
	dd H_2O	up to 25 ml

The buffer was aliquoted (100µl) and stored at –20^{o}C.

10 x running buffer:	Tris-HCl (pH 8.4)	30.3 g
	glycine	144.1 g
	SDS	10 g
	dd H_2O	up to 1000 ml

Coomassie staining buffer:

Coomassie blue R 250	1.5 g
methanol	455 ml
acetic acid	80 ml
dd H_2O	up to 1000 ml

Destaining solution:	methanol	50 ml
	acetic acid	70 ml
	dd H_2O	up to 1000 ml

SDS gels were set up and run in a minigel electrophoresis unit (mini-PROTEAN II; BioRad), using 7.3 cm x 10.2 cm glass plates and 0.75 mm spacers. The stacking and separating gel solutions (Table 5) were premixed without the addition of APS and TEMED. The polymerisation inducer (APS) and catalyst (TEMED) were added immediately before pouring the gels. After pouring the gels, a small volume of dd H_2O was overlaid onto the gel solution. After complete polymerisation the water layer was removed and the stacking gel was poured on the top of the separating gel. A plastic tenteeth comb was placed into the stacking gel. After stacking gel polymerisation the comb was removed, resulting in 10 wells used for applying the samples. The electrophoresis cell, containing one or two gel sandwiches was combined (see manufacturer's instructions), placed into an electrophoresis chamber and flooded with 1x running buffer.

Table 5. SDS-PAGE gel preparation

	Separating Gel (12 %)	**Stacking Gel (4%)**
dd H_2O	3.35 ml	3.05 ml
0.5 M Tris-HCl (pH 6.8)	2.5 ml	-
0.5 M Tris-HCl (pH 6.8)	-	1.25 ml
10 % (w/v) SDS	100µl	50 µl
30 % (w/v) Acrylamide / Bis	4 ml	665 µl
10 % (w/v) APS (fresh prepared)	50 µl	25 µl
TEMED	5 µl	5µl

The protein samples were mixed with ¼ volume sample buffer and incubated for 5 min at 100 ^{o}C. This step results in the denaturation of the proteins in the sample and complexing with SDS molecules. The samples were cooled on ice and equal protein amounts (8-10 µg protein per lane) were applied on the gel. The electrophoresis was run at constant current (25 mA / gel) until the bromphenol blue dye reached the end of the gel. The gel was then removed for Coomassie staining.

Coomassie staining was performed by agitating the gel in Coomassie staining solution for 1 hour at room temperature, followed by washing with destaining solution for at least 1 hour. The gels were documented by direct scanning and stored after drying with a BioRad GelAir Dryer.

The sizes of the analysed proteins were estimated by comparison with a standard protein marker mixture SDS-6H (Sigma) separated on the same gel.

3.2.2. Native PAGE

In contrast to the SDS-PAGE, in native PAGE the secondary, tertiary and quaternary structure of the proteins analysed is not affected. Here, the mobility of the proteins in the gel therefore depends on both their molecular weight and their charge under given buffer conditions.

Separating buffer (4x)	Tris-HCl pH 8.0	1.5	M
Stacking buffer (4x)	Tris-HCl pH 6.8	0.5	M
10x electrophoresis buffer	Tris	0.25	M
	Glycin	1.92	M
5x sample buffer	1M Tris-HCl pH 6.8	3.1	ml
	Glycerin	50	%
	Bromphenol blue	1	%
	dd H_2O	up to 10	ml

Table 6. Native PAGE gel preparation

	Separating Gel (7.5 %)	**Stacking Gel (3%)**
dd H_2O	1.8 ml	1.95 ml
30 % (w/v) Acrylamide / Bis	1 ml	0.3 ml
Separating buffer	1 ml	------
Stacking buffer	------	0.75 ml
TEMED	5 µl	5 µl
10 % (w/v) APS (fresh prepared)	40 µl	40 µl

The gel assembly, running conditions and Comassie staining and destaining procedures were the same as in the SDS-PAGE.

3.3. Concentration and dialysis of proteins

Proteins were concentrated using Amicon Ultra–15 ultrafiltration columns (Millipore) with 10 kDa cutoff. The columns permit over 60-fold concentration of the protein sample with a more than 90 % recovery.

Dialysis was used for desalting and/or buffer exchange, when handling bigger volumes of protein solutions and, in refolding experiments. The dialysis tubing used was VISKING Dialysis Tubing (Serva, Heidelberg) with a diameter of 16 mm. The tubing was boiled for 10 min in water and after filling with the protein containing solution was closed with clips. The dialysis was usually performed against a minimum of 50-fold excess of buffer overnight at 4°C.

3.4. Refolding of proteins

All protein refolding attempts applied in this work were preceded by purification of inclusion bodies from the *E. coli* host expressing the recombinant protein (see below) and subsequent denaturation of the protein sample in a buffer containing either 8 M urea or 6 M guanidine-HCl. A common protocol for inclusion body (IB) purification, which was used throughout this work is shown below (Vuillard *et al.*, 2002, modified).

- The pellet from 1 l of E: coli culture was resuspended in 20 ml of

50 mM Tris-HCl pH 7.5
0.5 M NaCl
1 mM PMSF
5 mM DTT
1 % v/v Triton X-100

The cells were opened by ultrasound sonication by bursts of 30 sec followed by cooling on ice until the solution clears. The IB were sedimented by centrifugation at 30,000 x *g* for 30 min at 4°C. The pellet (IB) was then washed twice with Tris pH 7.5 containing 1% Triton X-100 followed by spinning at 30,000 x *g* for 30 min at 4°C. The pellet (IB) was finally solubilised in 2 ml of:

50 mM Tris-HCl ph 7.5
6 M guanidine-HCl
25 mM DTT

and left for 1h at 4°C. Insoluble material was removed by centrifugation at 100,000 x *g* for 10 min. This final step is important to remove existing aggregates that can act as nuclei to trigger aggregation during refolding. The inclusion bodies purified with this method were further subjected to refolding experiments and the strategies applied are summarised below.

- Dialysis – the purified and solubilised (through denaturing) protein was subjected to step dialysis against different buffers with decreasing concentration of the denaturing agent, i.e. 8 M urea or 6 M guanidine-HCl. The steps used were – 4/2/0.5/0.1 M for urea and 3/1.5/0.5/0.1 M for guanidine-HCl. The buffers used were 50 mM sodium acetate and 50 mM sodium phosphate with pH ranging from 4 to 7.

- Rapid dilution – the protein sample was rapidly injected through a syringe needle into a buffer which contained no denaturing agent. The volume of the buffer exceeded the sample volume minimum 200-fold and the end protein concentration varied from 0.5 to 10 µg/ml in the different experiments. The injection of the sample was done stepwise (maximally 100 µl in each step when the buffer volume was 200 ml) while the buffer was vigorously stirred.

- *Vectrase* (BioVectra, Charlottetown, Canada) - the refolding experiments were done according to the manufacturer’s instructions. The method relies on a two-step process – a “capture” step, where a detergent forms a complex with the protein thus preventing its aggregation and a “stripping” step, in which the *Vectrase CD* reagent strips the detergent from the protein, allowing it to refold.

- FoldIt (Hampton Research, CA, USA). FoldIt is a factorial folding screening kit, which evaluates 12 factors on their ability to prevent protein aggregation and promote correct protein folding.

- Size exclusion chromatography (Harrowing *et al.*, 2003). The gel media of SEC allows simultaneous separation of denatured protein, folded protein and denaturant species while they pass through the column. Since the chance of protein–protein interaction is reduced during buffer exchange, the competing side reaction of aggregation is suppressed. Separation of the purified refolded protein from any aggregates that do form and buffer exchange is performed in one step.

3.5. Purification of proteins

3.5.1. Heat treatment of crude cellular extracts

The presumed thermostability of *P. torrridus* proteins permits the usage of heat treatment of *E. coli* cellular extracts containing the recombinant protein as a purification step. The temperature and time of heat treatment were optimised for each case. In general, the crude extract of a recombinant strain was subjected to heat treatment (50°C to 75 °C) at pH 6.5 for 10 to 20 min. The precipitated heat-labile proteins were pelleted by centrifugation (13 000 *g*, 20 min, 4°C) and the supernatant (heat treated fraction) was further processed for purification.

3.5.2. Fast Protein Liquid Chromatography (FPLC)

The FPLC techniques used in this work were done on an ÄKTA *design* system (Amersham Pharmacia Biotech, Uppsala, Sweden) equipped with a P-920 pump, UPC-900 monitor, INV-907 valve and a Frac-900 fraction collector. The protein separations were carried on HR 10/10 and XK 16/60 columns packed with Source 30Q and Superdex 200 material, respectively. The conditions used for the purification of two recombinantly expressed *P. torridus* proteins are listed in table 7.

Table 7. Conditions used for the purification of recombinant *P. torridus* glucose dehydrogenase (GdhA) and α-glucosidase (MalP)

		Anion exchange (Source Q30)	**Size exclusion** (Superdex 200)
GdhA	buffer A	50 mM Tris pH 8	50 mM Tris pH 8.0
	buffer B	A + 0.5 M NaCl	
	gradient	10 CV	isocratic, 1.2 CV fractionation
	flow rate	2 ml /min	1 ml/min
	fraction size	2 ml	5 ml
MalP	buffer A	50 mM Tris pH 9	50 mM Tris pH 9.0
	buffer B	A + 1 M NaCl	
	gradient	10 CV	isocratic, 0.8 CV fractionation
	flow rate	2 ml /min	1 ml/min
	fraction size	2 ml	3 ml

3.5.3. Purification of maltose binding protein (MBP) fusions

Fusion proteins containing the MBP as a partner were purified using the ability of MBP to bind maltose.

MBP buffer	20 mM Tris-HCl pH 7.4
	200 mM NaCl
Maltose	10 % w/v in MBP buffer
Amylose	10 % w/v in water

The amylose was subjected to three cycles of centrifugation (16,000 x *g*, RT) and resuspension in water in order to remove maximally the soluble amylose components and was mixed with 1.5 ml of MBP buffer. After two more cycles of centrifugation and resuspension the concentration of amylose was adjusted to 75 % and 50 µl of the obtained slurry was mixed with 50 µl cellular extract containing the MBP fusion protein. The mixture was incubated for 15 min on ice, centrifuged (16,000 x *g*, RT) and the proteins bound to the insoluble amylose were eluted with maltose solution.

3.6. Determination of enzyme activity and biochemical characterisation of enzymes

3.6.1. Glucose dehydrogenase (GdhA) enzyme activity measurements

Glucose dehydrogenase activity was routinely measured by spectrophotometrically monitoring (340 nm) the rate of release of NAD(P)H as a co-product of the reaction:

$$\text{Glucose} + \text{NAD(P)}^+ \rightarrow \text{Gluconolactone} + \text{NAD(P)H} + \text{H}^+$$

- The standard reaction was carried out in a volume of 1 ml at 55°C and contained:

		concentration in assay
phosphate buffer pH 6.5	970-x µl	50 mM
0.1 M NAD(P)$^+$	20 µl	2 mM
1 M D- glucose	50 µl	50 mM
enzyme	x µl	10 µg
total vol.	1000 µl	

The components were preheated for 10 min at 55°C and the reaction was started by the addition of glucose. Specific activity is expressed as µmol of NADPH produced per min per mg of protein under the specified conditions, using a molar absorptivity of 5,600 $M^{-1}cm^{-1}$ for NADPH. NAD$^+$ - dependent glucose dehydrogenase activity was measured the same way, substituting NAD$^+$ for NADP$^+$. Under the assay conditions, NADH had a molar absorption coefficient of 12,100 $M^{-1}cm^{-1}$.

An alternative method for measuring glucose dehydrogenase activity is to determine the rate of D-glucose decrease in the reaction. For these assays, the glucose decrease was assayed by the addition of 900µl freshly prepared enzyme-colour reagent solution to 100 µl of the reacted standard assay mixture (a commercially available glucose detection kit, Sigma 510-DA, was used). After 30 min of incubation at 37°C the glucose quantity was measured spectrophotometrically at λ = 450 nm.

For determination of the pH optimum of the enzyme (at 55°C, 10 min assay) and in pH stability testing, the following buffers were used:

50 mM glycine-HCl	pH 1.5 – 3.3
50 mM sodium acetate	pH 3.5 – 5.5
50 mM phosphate	pH 6.0 – 7.0
50 mM Tris HCl	pH 7.5 – 8.5

In these assays, the glucose dehydrogenase activity was measured by monitoring the decrease of D-glucose.

Glucose dehydrogenase activity was visualised on native PAGE gels by coupling the glucose dependent NADP$^+$ reduction to nitro blue tetrazolium formazan production (5- methyl phenazonium methyl sulphate, PMS, was used as an intermediate hydrogen carrier). For activity staining, the gel was soaked in 50 mM phosphate buffer containing 1mM NADP$^+$, 50 mM glucose, 1 mM NBT, 0.025 mM PMS for 10 – 15 min or until the appearance of a blue band. To normalise for the colour intensity, 30 mU glucose dehydrogenase were applied on each lane.

3.6.2. Alpha-glucosidase (MalP) enzyme activity measurements

Alpha-glucosidase activity was measured by using para–nitrophenyl α D-glucoside (pNPG) as a substrate. Enzyme cleavage of the arylglycosidic bond releases para-nitrophenol, which is quantified by measuring its absorbance at 420 nm. Under the standard assay conditions used for MalP, p-nitrophenol had a molar extinction coefficient of 12,000 M^{-1}cm^{-1}.

- standard pNPG reaction for MalP activity:

		concentration in assay
acetate buffer pH 4.5	90-x µl	50 mM
0.1 M pNPG	10 µl	10 mM
enzyme	x µl	0.3 µg
total vol.	100 µl	

The reaction was started by the addition of substrate and carried out at 85°C for 10 min. After the addition of 400 µl 1 M Na_2CO_3 and 500 µl dd H_20 the absorption was measured at 420 nm.

Enzyme kinetic analyses were performed by the direct linear plot method using Sigma plot v.8.0 (SPSS) for statistical evaluation of the data.

3.6.3. Thin layer chromatography (TLC) analysis of the α-glucosidase reaction products

This method permits the separation of low molecular weight compounds (mono- and oligosaccharides) based on their different mobility in a two-phase chromatographic system – a mobile (hydrophobic) and a stationary (hydrophilic) phase. As a stationary phase silica gel aluminium plates were used (Silica gel 60 F_{254}, Merck, Darmstadt, Germany). The mobile phase consisted of:

Running solution

1- propanol	50 ml
nitromethan	30 ml
H2O$_{bidest}$.	20 ml
Running time	2h

Developing reagent

anilin	1 ml
diphenylamin	1 g
aceton	100 ml

The test reaction was carried out as described in chapter B.3.6.2. and a 2 µl aliquot was applied on the TLC plate. After carrying out the separation in a sealed chamber, saturated with 100 ml of the running solution, the TLC plate was left to dry, sprayed with freshly prepared spray solution (developing reagent + 1/10 vol. phosphoric acid) and baked in an aluminium foil for 12 min at 140°C.

C. RESULTS

1. Sequencing, assembly and sequence analysis of the *P. torridus* genome

1.1. Sequencing, assembly and editing

The aim in the sequencing and assembly phase of a genome sequencing project is to generate a sufficient number of sequences which, after assembly, yield a complete and redundant coverage of the genome. The number of sequences necessary to obtain a given coverage of a genome of a certain size can be simply calculated using the formula:

$$N = \frac{A * L}{w} \qquad (1)$$

where *N* denotes the number of sequences, *w* their average length, *A* the wished coverage of the genome and *L* the genome size. In the case of the *P. torridus* genome, setting a genome size of 1.5 Mb, an average sequencing length of 600 bp and an 8-fold coverage of the genome, 20,000 single sequences or 10,000 shotgun library clones would be needed.

Statistical methods exist with whose help it is possible to model the relationship between the number of sequences (of a certain size) and the number of contigs (respectively gaps) that would be obtained, using the shotgun approach (Lander and Waterman, 1988) :

$$P = e^{\frac{-N * w}{L}} \qquad (2)$$

N, L and *w* are as in the previous formula, and *P* describes the probability of not covering a given base pair position of the genome. Using the above values for the *P. torridus* genome, it can be calculated that with these parameters the resulting *P* value is 0.03 %, meaning that 450 bp would not be covered in a 1.5 Mb genome. Assuming an average gap size of 75 bp the sequence would be comprised of 6 contigs. These theoretical considerations are based on the assumption that all parts of the genome can be sequenced equally well and do not take into account the quality of the sequence reads.

A total of 24,224 sequences, generated at Integrated Genomics, were assembled with the use of the PHRAP assembly tool (Ewing *et al.*, 1998) leading to a database consisting of 26 contigs. The sequences along the contigs were equally distributed, *i.e.* without highly overrepresented regions. This high quality data was further edited with the help of the Staden software package. Editing with the program gap4, implemented in the Staden package, was facilitated by the use of *phred* confidence values (Bonfield *et al.*, 1995). These values are used to calculate the confidence of the consensus sequence (the probability of an error) based on the quality of the sequence reads and hence easily identify places requiring visual trace inspection or extra data. Sequence analysis and editing was done in close collaboration with Dr. Ole Fütterer.

1.2. Gap closure and further editing

Closing of the genome gaps was accomplished primarily by primer walking – designing primers from the ends of the contigs and using them for additional sequencing reactions.

Another useful method which was used for closing gaps was to compare the ends of the contigs with the help of BLAST searches with already sequenced genomes of related organisms. This approach is particularly applicable when the contig ends contain conservative regions in which the gene order is preserved in related organisms or are within coding regions which have close homologs. An example of such a case in the *P. torridus* genome is shown on Figure 2.

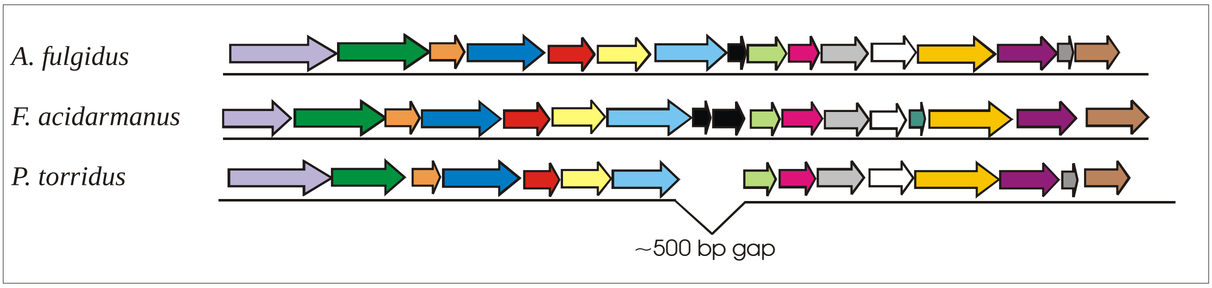


Figure 2. BLASTP comparison of the ends of two *P. torridus* contig ends and the genomes of *Archaeoglobus fulgidus* and *Ferroplasma acidarmanus*. The ORFs marked with the same colour are orthologs. The gene cluster corresponds to ribosomal subunit genes.

Each pair of contig ends that was inferred by this method to be neighbouring was confirmed by PCR on genomic DNA as a template and the PCR product sequenced. The trace obtained was introduced in the sequence database to verify its consistence and used for closing of gaps. In addition, uncertain sequence regions identified during the editing process were re-sequenced on the respective shotgun library plasmids using vector specific primers.

An additional approach, implemented in the closing of gaps was multiplex combinatorial PCR (Tettelin *et al.*, 1999). Briefly, the primers flanking the ends of the contigs obtained after shotgun assembly (26 contigs and 52 primers) were combined in 9 pools (8 pools of 6 primers and 1 pool of 4 primers), and combinatorial PCR reactions were performed using *P. torridus* genomic DNA as a template and *Pfu* DNA polymerase in which each primer pool was combined with all of the other pools. Thus, the number of the reactions necessary to obtain full combining of the pools is: $\binom{9}{2} = 9 * 8 / 2 = 36$. After analysing the PCR products by agarose gel electrophoresis the cases in which a single PCR product was obtained were directly submitted to sequencing using each primer pool in a separate sequencing reaction and the PCR product as template (Fig. 3). In the rest of the cases, additional combinatorial PCR reactions were necessary in order to identify the primer pairs that produced the PCR product.

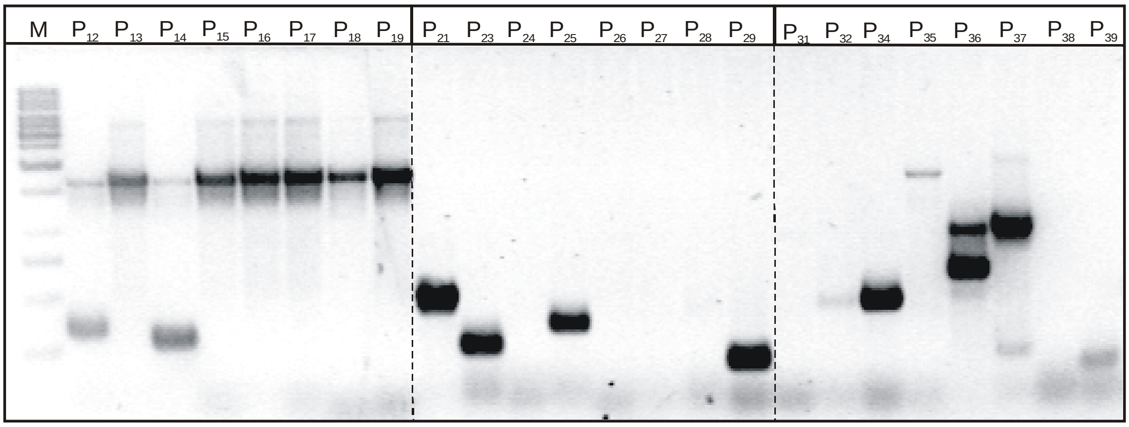


Fig. 3. Agarose gel electrophoresis (1.5 % agarose) showing multiplex combinatorial PCR (MCPCR) products of the full combining of three of the primer pools (1 to 3). It can be seen that the primer pool 1 contains an intrapool primer pair: there is one PCR product (about 1.6 kb) present in all the combinations. M: 1kb DNA marker, P_{xx}: pool combination.

A total of 1,470 sequence reads were necessary to completely cover the *P. torridus* genome during this editing phase, using the above methods. It is worth noting that no inconsistencies were observed between the results obtained by the different gap closing approaches.

1.3. Sequence analysis

The closed *P. torridus* genome represents a 1,545,900 base pairs large single circular chromosome. The overall number of sequence reads was 25,694 which yielded a 9.4-fold coverage of the genome and a statistical error rate of below 1 in 2,000,000.
Sequence analysis of the *P. torridus* genome was performed using different software tools, as described in section B.2.6. After ORF prediction, the sequence information was implemented in the annotation program ERGO (Overbeek *et al.*, 2000 Integrated Genomics). The program predicts functions for all the ORFs based on their amino acid sequence similarity to ORFs present in the internal non-redundant ERGO database. Each automatically annotated ORF was manually inspected and the annotation confirmed. Annotation of the ORFs of the P. torridus genome project was done in close collaboration with Dr. Ole Fütterer (Applied Microbiology, Georg-August-Universität, Göttingen). In addition, about 30 % of the ORFs were annotated by our project partners at the Technical University Hamburg-Harburg (Prof. G. Antranikian). Further, metabolic pathway reconstruction was accomplished using the information obtained in the ORF annotation phase.

1.3.1. General features of the *P. torridus* genome

For 64 % of all ORFs found in the genome it was possible to assign a function. Of the remaining 553 hypothetical ORFs 73 are unique to *P. torridus* while 480 showed similarities to hypothetical ORFs of other organisms (Table 8). A prominent feature of the genome is its high coding density: 91.7 % of the DNA was found to be coding sequence. Some general characteristics of the *P. torridus* genome are presented in Table 8. The spatial distribution and the functional grouping of the predicted ORFs on the circular map of the *P. torridus* genome is depicted in Fig. 4 together with the repeats found and the GC content deviation.

Table 8. General features of the *P. torridus* genome

General features	**Number**
Size (bp)	**1,545,900**
Coding region (%)	**91,7**
G + C content (%)	**36**
Total number of open reading frames	**1535**
ORFs with assigned function	**982**
ORFs without function	**553**
ORFs without function or similarity	**73**
ORFs without function, with similarity	**480**
ORFs with putative signal peptides	**121**
ORFs involved in transport	**170**
Stable RNAs	
rRNAs	**1 5S; 1 16S ; 1 23S**
tRNAs	**47**
IS-elements	**4**

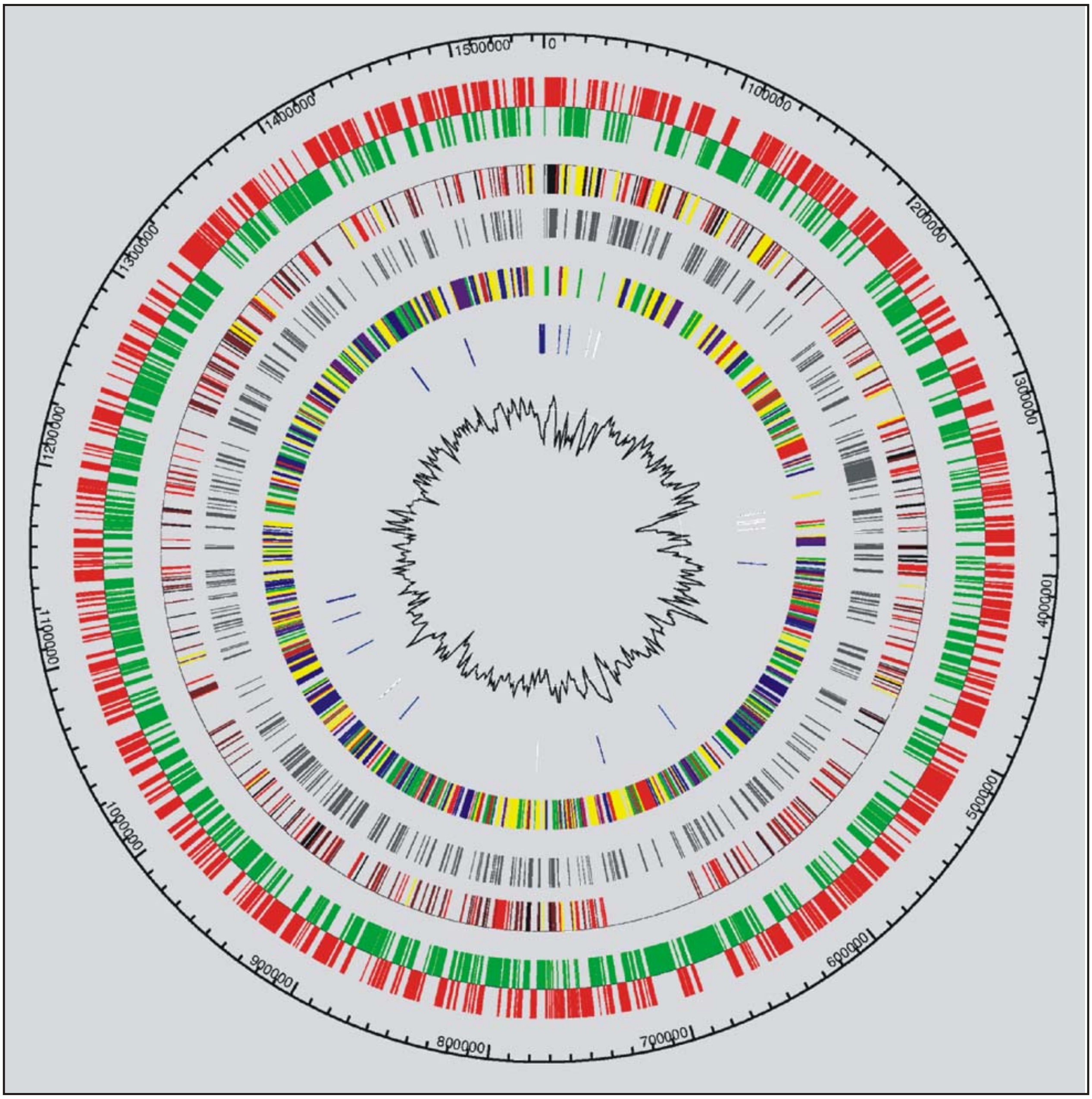


Figure 4. Circular map of the single chromosome of *P. torridus*.
Cirles from outside to inside: (1) orf orientation. Red = + strand, green = - strand. (2) ORF origin. Red = archaeal, yellow = bacterial, grey = thermoacidophilic, black = unique. (3) Hypothetical ORFs. (4) Functional categories. Red = protein synthesis, yellow = transport, light blue= energy metabolism, light green = nucleotide metabolism, light brown = DNA metabolism, turqois = RNA metabolism, grey = hypothetical, black = disordered, other = white. (5) Lenght of direct repeats. Light blue = < 50 bp, blue = < 100bp, dark blue < 200bp, black = < 500bp. (6) GC content deviation. (%GC-content in 2500b)-(average GC-content all)

1.3.2. Comparative analysis of the amino acid composition and the isoelectric point distribution of *P. torridus* proteins

These analyses were performed by comparing the whole genome amino acid composition and isoelectric point distribution of the *P. torridus* predicted proteins with those of organisms having different degree of phylogenetic relatedness (using the data available in the internal ERGO database as of 12.2003). These results are presented in Figures 5 and 6.

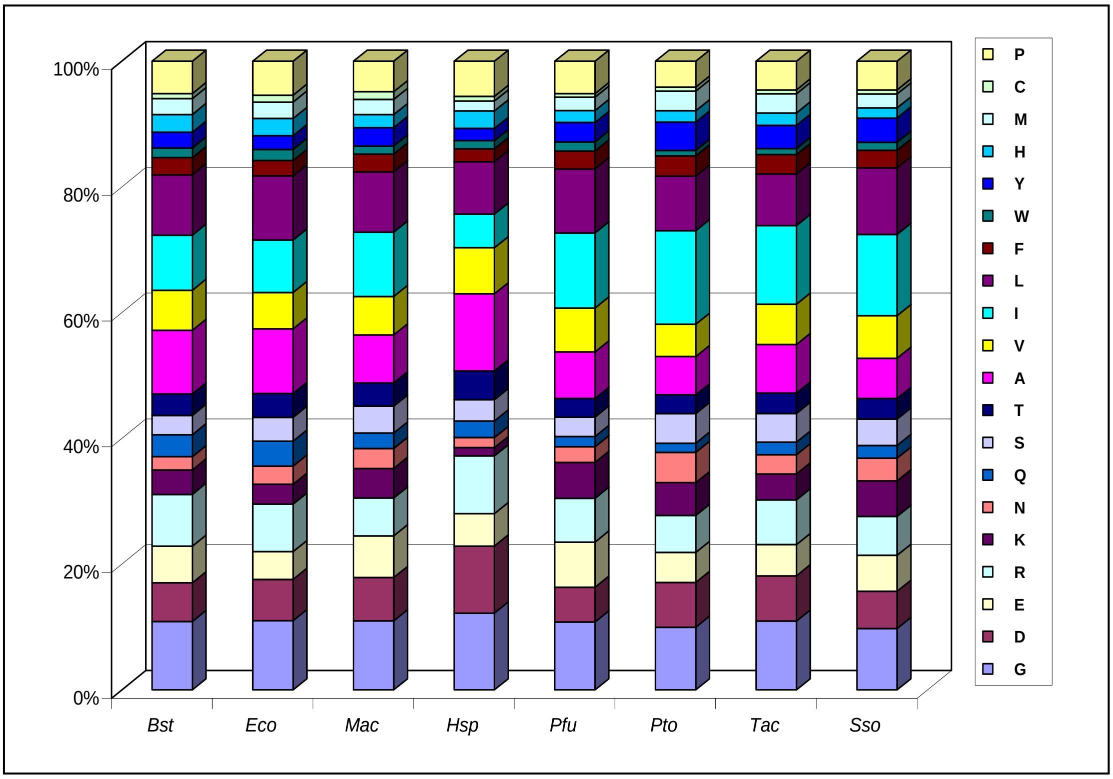


Figure 5. Comparison of the whole genome amino acid composition in: *Bst - Bacillus stearothermophilus, Eco – Escherichia coli, Mac - Methanosarcina acetivorans, Hsp - Halobacterium sp. NRC1, Pfu - Pyrococcus furiosus, Pto – P. torridus, Tac - Thermoplasma acidophilum, Sso - Sulfolobus solfataricus*. The amino acid color designations are shown on the right.

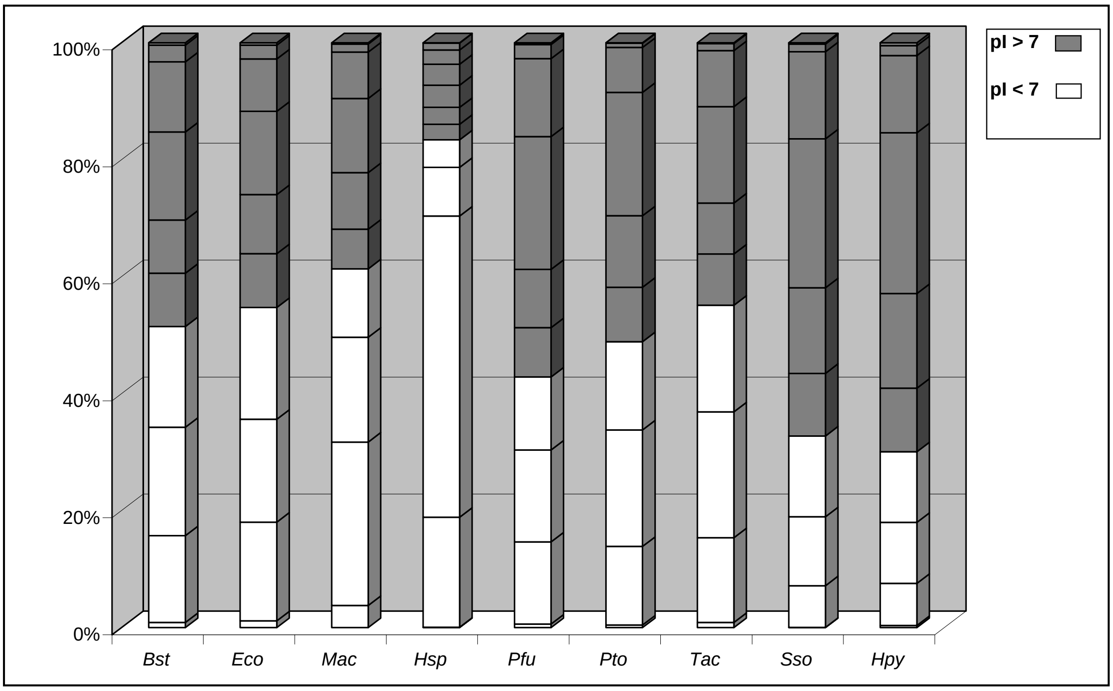


Figure 6. Comparison of the whole genome isoelectric point distribution in *Bst - Bacillus stearothermophilus, Eco – Escherichia coli, Mac - Methanosarcina acetivorans, Hsp - Halobacterium sp.* NRC1, *Pfu - Pyrococcus furiosus, Pto – P. torridus, Tac - Thermoplasma acidophilum, Sso - Sulfolobus solfataricus, Hpy – Helicobacter pylori.* Each block represents a pI range of one unit.

1.3.2. Origin of replication

The origin of replication of the *P. torridus* chromosome could be identified based on keto- and purine-excess plots (Figure 2 and Appendix A), the gene contents of this genome region as well as on the clustering of repeats in this region. *P. torridus* contains several long direct and inverted repeats in the range of 100 – 369 bp (Table 9). While all long repeats lie in coding regions and have only 2 copies, 24 short direct repeats with a length of 35-39 bp are found clustered in a non-coding stretch of 3000 bp which at the same time represents the largest non-coding region in the genome. Keto- and purine-excess plots localise the origin of replication in this area and coding regions for a DNA-polymerase and a DNA-helicase are found close to its 3’ end. Although no inverted repeats can be found and the gene for the Cdc6/Orc1-homolog is not in close proximity, we assume this region to be the origin of replication in this organism.

Table 9. Repeated sequences found in the genome of *P. torridus*

Repeats	Number	Remarks
Direct repeats		
Long direct repeats 100 – 317 bp	10	2 copies, 1 near assigned OriR
Short direct repeats 40 – 100 bp	15	2 copies, 4 near assigned OriR
Short direct repeats 35 – 39 bp	29	2-4 copies, 24 near assigned OriR
Inverted repeats		
Long inverted repeats 100 – 369 bp	4	2 copies
Short inverted repeats 35 - 100 bp	21	2 copies

In the neighbourhood of the predicted oriR many genes were found which are unique to *P. torridus* as well as a high concentration of genes which so far were only found in *Bacteria*.

1.3.3. Replication, repair and restriction

The replication apparatus of *P. torridus* is of the classical archaeal type. The ORF for an Orc1/Cdc6 homolog which recognizes the replication origin was found near the ORF for a DNA-helicase probably involved in the unwinding of the parental duplex DNA in cooperation with single-strand DNA binding proteins and a topoisomerase. Coding sequences for a two-subunit gyrase were found in the genome whereas no reverse gyrase gene could be detected. Synthesis of the RNA/DNA primer can be accomplished by a two-subunit primase and genes coding for all of the DNA-polymerase complex proteins necessary for strand elongation were identified: ‘clamp loader’ and ‘sliding clamp’ polypeptides as well as three DNA-polymerases of the DNA-polymerase families X, B and D. An endonuclease and a ribonuclease were identified which could remove RNA primers attached to the 5’-end of the Okazaki-fragments before gap-filling and joining by a ligase.

In order to ensure DNA integrity, *P. torridus* contains the coding capacity for a large number of repair and recombination proteins. Two repair endonucleases of type III, and of type IV, and one of type V, three repair DNA-helicases, two proteins with

MutT-like domains and the repair proteins RadA, RadB, MRE11, Rad50 are exclusively involved in DNA repair or, in part, play a role in recombination processes, together with a RecJ exonuclease homolog and the topoisomerase and ligase already mentioned. Besides a type II restriction/modification system usually found in the genomes of other thermoacidophiles, *P. torridus* also posesses a type I system.

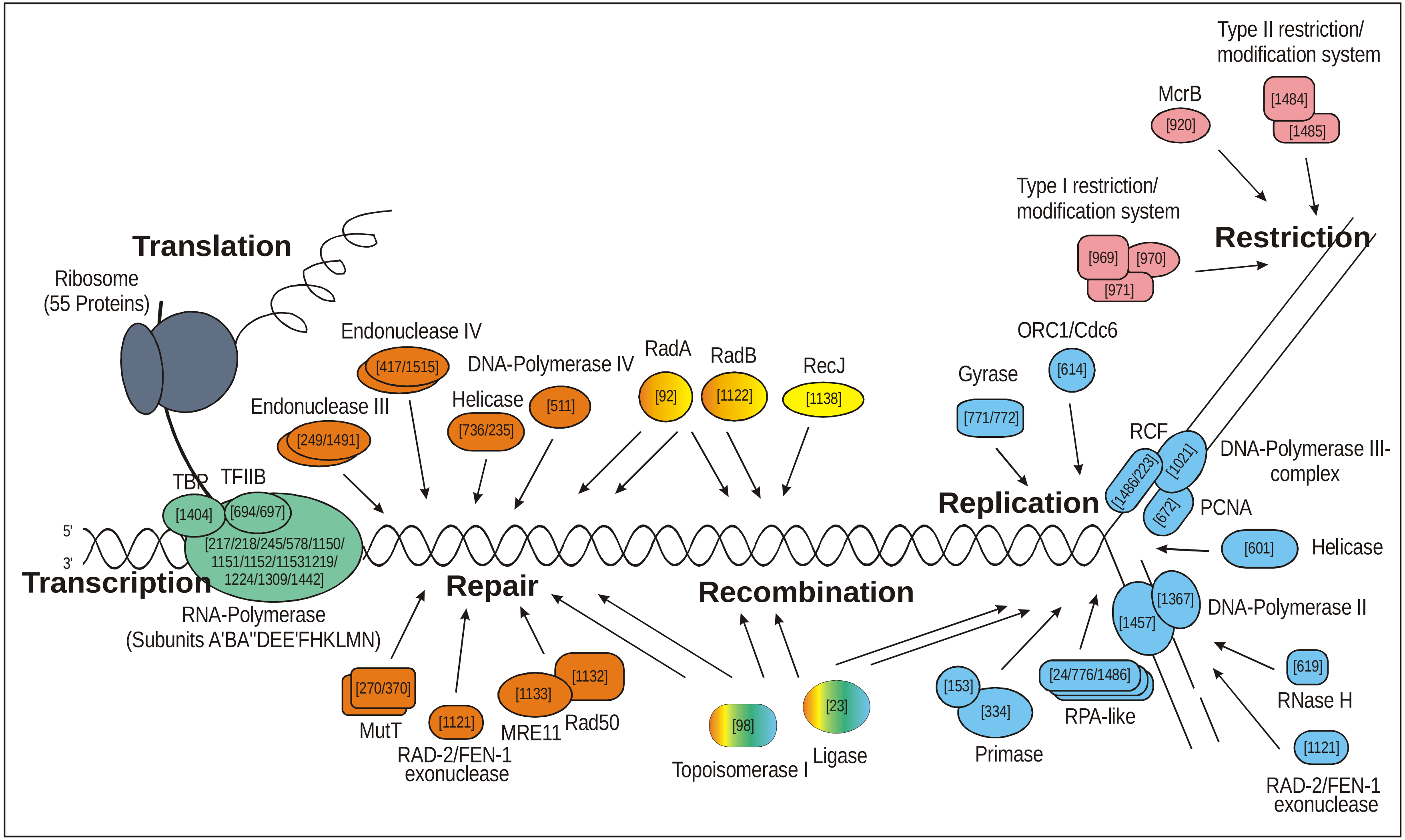


Figure 7. Overview of the replication, transcription and translation, recombination and repair proteins, identified in the P. torridus genome. The ORF numbers are in square brackets and are as in the ERGO database as of 12.2003.

1.3.4. Amino acid metabolism

Biosynthetic pathways for all 20 amino acids were reconstructed in the *P. torridus* genome. The anabolic histidine gene cluster is the second largest cluster of functionally connected ORFs in the genome after the main ribosomal cluster. It has its highest similarities in nucleotide sequence and operon structure to the corresponding homologs of *F. acidarmanus*, members of the *Pyrococcales* and different bacteria and clusters with them in protein phylogenetic trees (data not shown).

With respect to the utilisation of amino acids, a major source of carbon and energy for *Picrophilus*, it has been reported that *P. oshimae* cells rapidly take up the amino acids histidine, proline, glutamate and serine, although only glutamate, proline and leucine were able to drive respiration as was determined via monitoring changes in the external pH (van de Vossenberg *et al.*, 1998). Genome analysis of *P. torridus* revealed that this organism possesses in particular genes and pathways for the degradation of aspartate, glutamate, serine, arginine, histidine, glycin, threonine and the aromatic amino acids phenylalanine and tyrosine. In some of the pathways energy is generated either by substrate-level or oxidative phosphorylation. Arginine is catabolised via the arginine deiminase and subsequent ornithine/citrulline cycle. ATP is generated directly by the dephosphorylation of carbamoylphosphate as the final step in this pathway which is used by other thermoacidophilic archaea and many strictly anaerobic bacteria. For histidine degradation, the genes required for the reaction from histidine to urocanate and for the subsequent conversions into glutamate and formiminotetrahydrofolate (by glutamate formiminotransferase) or glutamate and formamide (by formiminoglutamase) were found. Subsequently, a predicted formamidase could convert the latter into formate and ammonia. Interestingly, so far, both pathways are detected only in the thermoacidophilic members of the archaea. Glutamate could be metabolised to 2-oxoglutarate by a predicted glutamate dehydrogenase but genes are present also for a glutamate decarboxylase, an aminobutyrate aminotransferase and a succinate-semialdehyde dehydrogenase eventually yielding succinate. Products of both pathways can be further metabolised in the TCA cycle.

The breakdown of serine, glycine and histidine in *P. torridus* requires an operating folate or modified folate C1-metabolism. Some of the genes for later steps in tetrahydrofolate (THF) biosynthesis were found as well as all genes needed for the backbone of the one-carbon transfer reactions, from THF and formate to 5,10-methylene THF. The availability of a one-carbon folate pool can greatly enhance the metabolic capacity of an organism as it can be used to gain reducing equivalents by various catabolic reactions and to provide C1-compounds for nucleotide, methionine and pantothenate biosynthesis. In *P. torridus*, most of the genes at the periphery of the one-carbon folate pool have catabolic functions apart from some involved in purine and methionine biosynthesis. No THF-dependent genes have been found for the synthesis of pantothenate, formylmethionine-tRNA or thymidine. In most archaea, C1-compounds are carried by modified pterin-containing compounds which are structurally related to folate (Angelaccio *et al.*, 2003). Since the relationship of folate or folate derivative-dependent enzymes is very high, it is unclear what compound is used in *P. torridus*.

1.3.5. Protein and peptide degradation

As *P. torridus* is believed to live as a scavenger (Schleper *et al.*, 1995), peptides and proteins are important growth substrates. Proteins can be degraded by several predicted extracellular acid proteases, including two thermopsin-like proteins, and two serine proteases. Two of the predicted proteases (ORFs 844 and 1054, annotated as thermopsin precursors) were cloned into expression vectors (see section 2). Most of these proteins possess a putative transmembrane helix at their C-terminal end which is thought to serve as a membrane anchor. This observation can also be made when looking at exported proteins from many other archaea (Albers *et al.*, 2001) and also seems to be true for other extracellular proteins of *P. torridus*. Four ABC-transporters were found responsible for the uptake of di- and oligopeptides which can be further degraded to free amino acids by a tricorn petidase, two tricorn cofactors (F2 and F3), an acylaminoacyl peptidase, a proline dipeptidase and a metallo-carboxypeptidase.

1.3.6. Protein synthesis and export

A large number of ORFs predicted to encode chaperones were found: the Hsp60 system (two thermosome subunits and two prefoldin/Gim subunits), a VAT-protein, a

Lon-2-related ATPase, two Hsp20 proteins and also the complete Hsp70 system DnaK, DnaJ and GrpE which is absent from most other archaea (Ruepp *et al.*, 2000). Genes required for the twin arginine and signal recognition protein export systems could be detected, and a total of 121 putatively exported proteins were predicted by the SIGNALP-algorithm (Appendix D). Most of them were annotated as transporters, exported binding proteins, proteases, components of the respiratory chain or hypothetical proteins. 38 % of the predicted exported proteins possess a C-terminal transmembrane helix which could serve as a membrane anchor. Interestingly, in five genes annotated as ABC-transport binding proteins signal peptides could be detected but no means for anchoring the proteins to the cell wall or membrane.

1.3.7. Transporters

The genome of *P. torridus* contains a large number of genes coding for transporters. 170 ORFs or 12 % of all genes play a role in transport. 21 transporters are predicted to be involved in drug export. It was assumed that most of these are required in detoxification of the cell as we could not detect any genes for secondary metabolite biosynthesis. Uptake systems for Fe^{3+}, NH_4^+, Cu^{2+}, Mn^{2+}, Zn^{2+}, SO_4^{2-} and phosphate were found, as well as two proton/sodium exchangers and transport channels for Cl^- and K^+. Also present are several transporters for nucleotides and a number of organic acids. A large number of ORFs seem to be necessary for the uptake of peptides, amino acids (34) and sugars (32). Nearly half of them code for ABC-transporter subunits (Fig. 8).

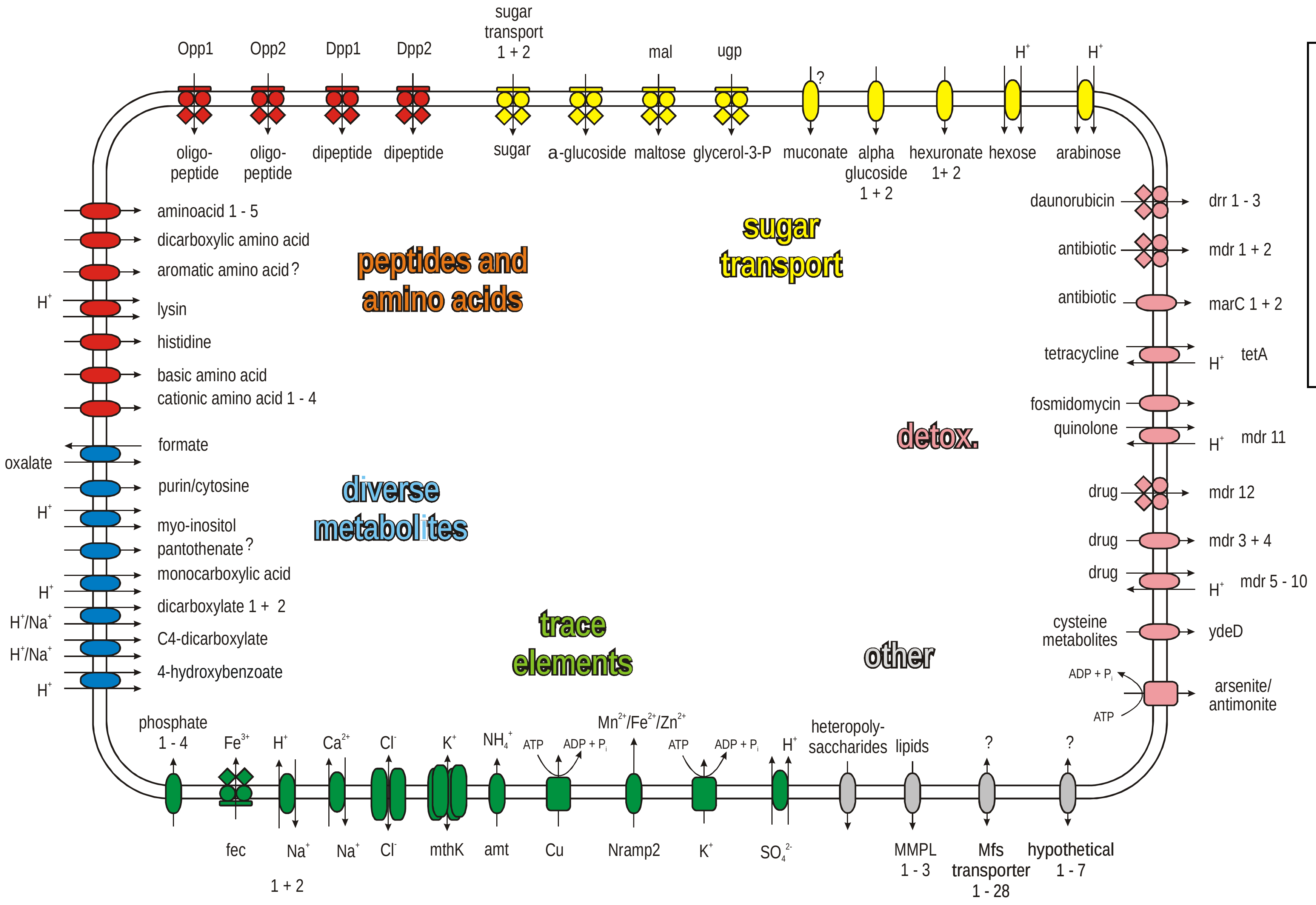


Figure 8. Overview of the transporters identified in the genome of *P. torridus.*

secondary transporters

primary (ATP dependent) transporters

1.3.8. Energy metabolism

In the acidic and hot environment of *P. torridus* the polymeric sugar compounds outside of the cell are probably subject to non-enzymatic degradation as well as to enzymatic attack. An extremely acid-stable extracellular glucoamylase of *Picrophilus* has been recently investigated (Serour *et al.*, 2003). Genes for five ABC sugar uptake systems and seven secondary transporters were identified whose products can transport oligomeric and monomeric sugar molecules into the cell. Two predicted intracellular glucoamylases, an α-glucosidase and an α-amylase can further degrade oligomers to glucose. The predicted α-glucosidase (ORF 985) was successfully expressed in *E. coli* and its annotation was confirmed experimentally (section C.3.1.2).

P. torridus most likely catabolises glucose via a non-phosphorylated variant of the Entner-Doudoroff (ED) pathway. Genes for all steps have been assigned including a gluconate dehydratase gene, although no annotated sequence data is available for this ORF in public databases. Moreover, $NADP^+$-dependent glucose dehydrogenase activity in *P. torridus* cells could be detected and the gene coding for this activity was successfully expressed in *E. coli* (see section C.2.2). The Embden-Meyerhof-Parnas pathway is assumed also to be functional in *P. torridus* despite of the lack of a gene with similarity to both classical types (I and II) and also to the archaeal type IA fructose-1,6-bisphosphate aldolase for reasons which are discussed later (see discussion section D.4).

None of the enzymes that catalyse the conversion of glucose into 6-phosphogluconate could be identified suggesting that the pentose-phosphate pathway is not used for glucose catabolism in *P. torridus*. However, a transketolase, a transaldolase, and both genes required for the interconversion of the pentose phosphates are present to provide the cell with nucleotide precursor molecules.

Glycogen-like intracellular polysaccharides have been detected in thermoacidophilic archaea, i.e. *T. acidophilum* and *S. solfataricus* (König *et al.*, 1982) but it is still unclear whether *P. torridus* is able to synthesise storage polysaccharides. Most biosynthetic genes were identified in the genome sequence, i.e. those coding for a phosphoglucomutase and a glucose-1-phosphate uridylyltransferase (UDP glucose pyrophosphorylase). However, neither *P. torridus* nor *T. acidophilum* appear to contain a classical glycogen synthase homolog. For the catabolic pathway, a putative glycogen debranching enzyme was found clustered with a family 57 α-amylase which is thought to be involved in the breakdown of intracellular storage compounds.

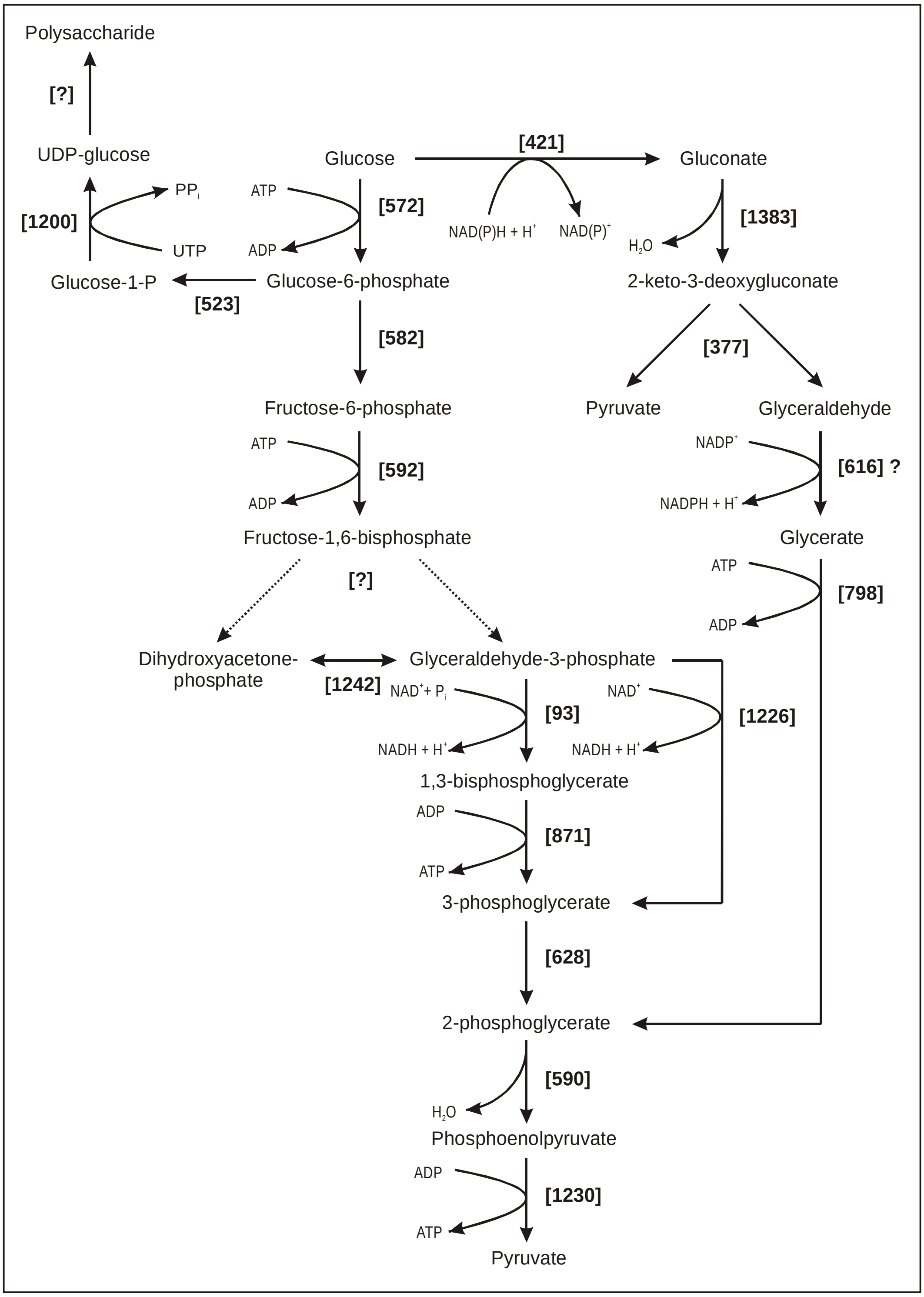


Figure 9. Reconstruction of the Embden-Meyerhof-Parnas and Entner-Doudoroff pathways of *P. torridus*. ORF numbers are in square brackets and question marks denote the lack of a gene for the respective step or uncertainty in the annotation.

Pyruvate as the final product of glycolysis can be converted to acetyl-CoA by either a NAD$^+$-dependent pyruvate dehydrogenase or a ferredoxin-dependent pyruvate oxidoreductase. It is unclear at present if both enzymes operate *in vivo*. Also, it is not known if the reaction is used in the reverse direction for anabolic purposes. The homolog of the ferredoxin-dependent enzyme from the autotrophic archaeon *Methanobacterium thermoautotrophicum* is thought to be able to catalyse the reverse reaction (Tersteegen *et al.*, 1997) from acetyl-CoA to pyruvate assimilating CO_2 in this step.

P. torridus appears to contain a complete set of genes for the oxidative tricarboxylic acid cycle (TCA) for the oxidation of acetyl-CoA. In parallel, the organism maintains the coding capacity for the 2-methylcitrate pathway for the oxidation of propionyl-CoA using the enzymes of the TCA-cycle responsible for the conversion of succinate to oxaloacetate. A gene coding for a propionyl-CoA synthase has been detected enabling *P. torridus* to grow on propionate.

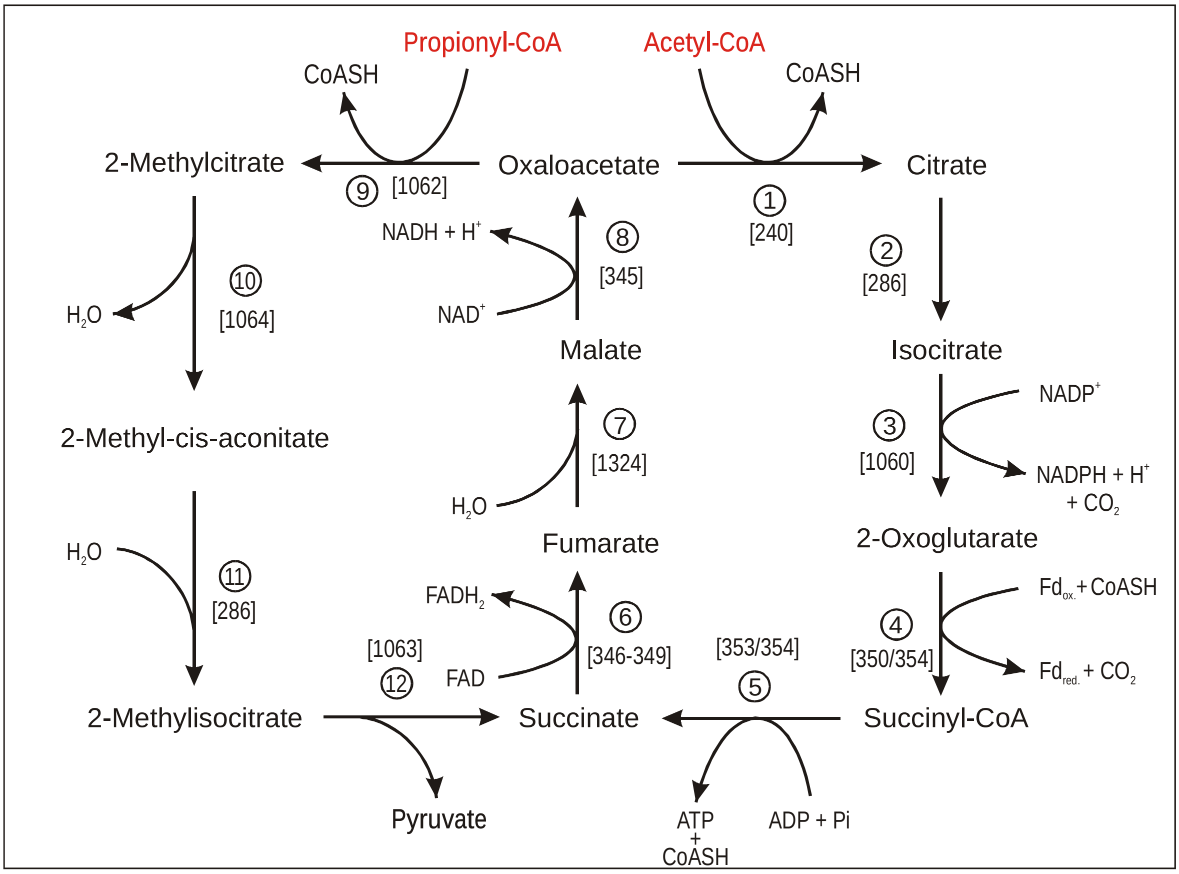


Figure 10. Tricarboxylic acid and 2-methylcitrate cycles reconstruction of *P. torridus*. The numbers in circles indicate the sequence of the cycle reactions.

In contrast to its close relative, the microaerophilic *T. acidophilum, P. torridus* is an obligate aerobe and apparently uses a more complex electron transport chain in order to generate the membrane potential (Fig. 11). Although no complete set of quinone synthesis genes were identified, many genes for protein complexes were found which feed their electrons into the electron transport chain, among them several quinone oxidoreductases. Among the latter are a pyruvate oxidase, a CO-dehydrogenase, a formate lyase complex and a sulfide-quinone oxidoreductase. Another gene was found coding for a putative sulfide dehydrogenase which transfers the electrons directly to cytochromes.

Quinol oxidation in *P. torridus* is accomplished by a complex similar to the SoxM-complex which has been desribed in *S. acidocaldarius* (Lubben *et al.*, 1994). This complex consists of the quinone- and the terminal-oxidase with a blue copper protein (sulfocyanin) as electron shuttle between the two. While homologs of the *P. torridus* quinone oxidase were found in the *T. acidophilum* and *S. solfataricus* genomes, significant similarities of the *P. torridus* terminal oxidase part were only found with bacterial cytochrom c oxidases. Genes similar to the blue copper protein gene of *P. torridus* were only found in the *Sulfolobus* and *Ferroplasma* genomes.

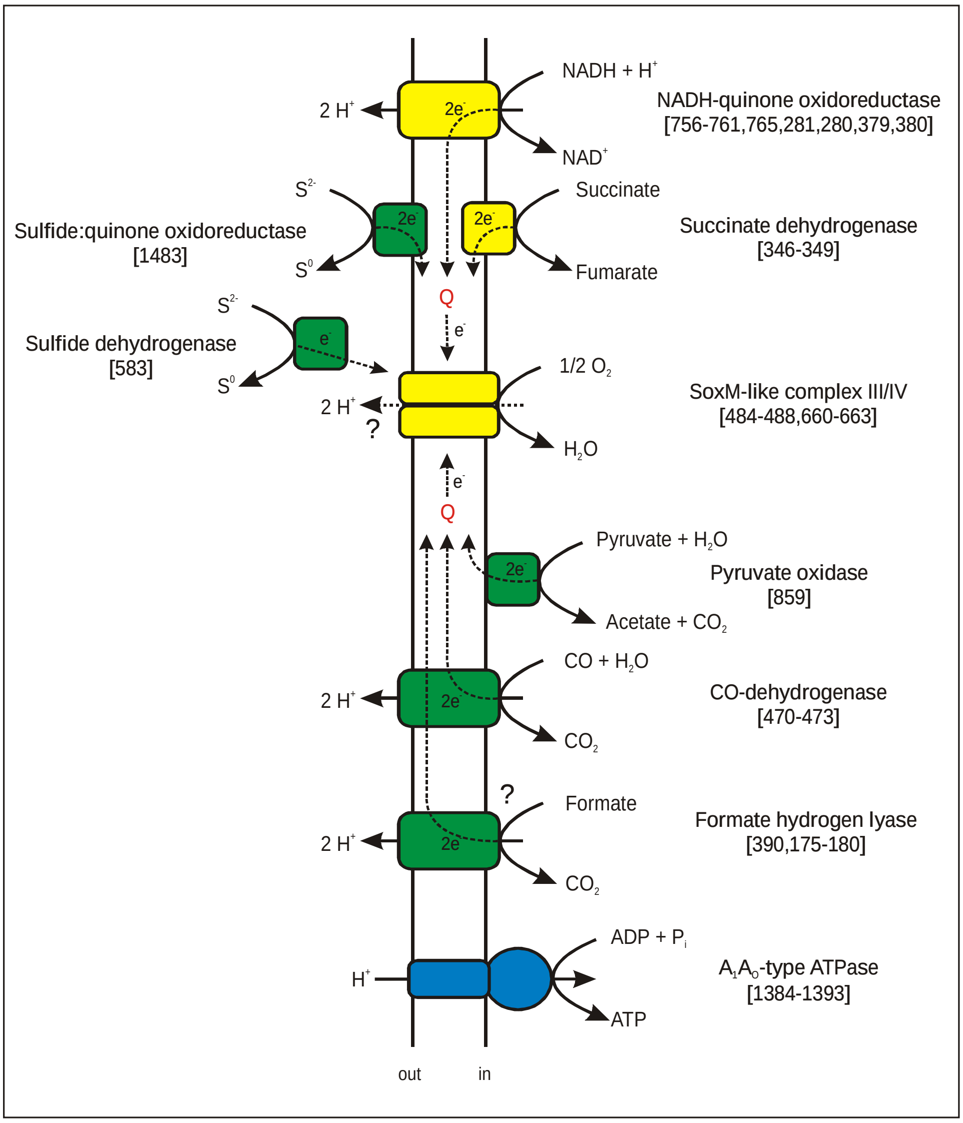


Figure 11. Genes coding for respiratory chain components as inferred from genome data.

1.3.9. Porphyrin metabolism

P. torridus seems to be able to synthesize porphyrins like cytochromes and adenosylcobalamin. In the annotated genome, 28 genes were found involved in the synthesis of adenosyl-cobalamin from glutamate, which represents nearly two percent of the whole genome. This anabolic pathway starts with L-glutamate and proceeds via glutamate-1-semialdehyde and aminolevulinate to uroporphyrinogen III, the last common precursor of the porphyrins. For cobalamin biosynthesis, uroporphyrinogen III is converted to precorrin 2 in which either Co^{2+} is inserted in organisms employing the anaerobic route yielding cobalt-precorrin 2, or which undergoes a methylation step followed by an oxygen-dependent ring contraction yielding precorrin 3 in the aerobic pathway. Organisms using the anaerobic pathway are known to possess type II, or ‘early’ metal ion chelatases for cobalt insertion (Scott *et al.*, 2002). *P. torridus* contains no type II metal chelatase-homologous genes and, in contrast to *Thermoplasma* and *Sulfolobus*, seems to employ the aerobic pathway where the insertion of the cobalt ion takes place after the main modifications of the porphyrin ring system and is catalysed by a type I or ‘late’ metal ion chelatase protein complex similar to CobNST of *Pseudomonas denitrificans*. However, most enzymes of the porphyrin biosynthesis of *P. torridus* revealed their highest similarity scores with sequences deduced from the porphyrin biosynthesis genes of other thermoacidophilic archaea which employ the anaerobic pathway, thus indicating that the ‘late’ metal ion chelatase complex was acquired later through horizontal gene transfer.

Interestingly, *cob*S or *cob*T homologs could not be found in the *P. torridus* genome. Instead, two genes homologous to magnesium chelatase subunits *chl*I and *chl*D flanking the *cob*N gene could be identified. Comparative studies with other prokaryotes (data not shown) showed that this is not uncommon, and it has recently been suggested that ChlI and ChlD may take over the function of CobS and CobT (Rodionov *et al.*, 2003).

1.3.10. Oxigen stress genes

As a strict aerobic organism, *P. torridus* possesses several mechanisms to protect the cell against oxidative damage. Genes coding for a superoxide dismutase, three putative peroxiredoxin-like proteins and an alkyl hydroperoxide reductase were found, the last is present only in thermoacidophilic archaea and *P. furiosus*. Flanking the predicted OriR region a β-carotene biosynthetic operon strongly resembling genes from marine ε-proteobacteria and corynebacteria was detected. B-carotene formation in other archaea has only been predicted for *S. solfataricus* and biochemically studied in *Halobacteria* (Hemmi *et al*., 2003 and Spudich *et al*., 2000).

2. Heterologous expression of *P. torridus* genes

In order to further investigate the adaptation mechanisms, allowing *P. torridus* to survive at low pH values, a part of the current project was to attempt cloning and heterologous expression of selected genes from this organism. As synthesis of heterologous proteins in *E. coli* is a well developed, economical and convenient method, it was chosen as a common approach. A list of *P. torridus* ORFs that were cloned in *E. coli* with the purpose of overproducing the encoded proteins is shown in Table 10. Of these, ORFs 985 (α-glucosidase) and 421 (glucose-1-dehydrogenase) were successfully expressed and the purification and biochemical characterisation of the encoded enzymes is described in chapter 3.

Table 10. *P. torridus* ORFs cloned in either pCR4_TOPO or pDrive vectors. The ORFs in bold were further subcloned into different expression vectors (described below). Shaded ORFs were expressed as functional enzymes. The ORF numbers are as in the ERGO database (as of 04.2004)

ORF	Predicted function	Construct
615	**β-galactosidase (EC 3.2.1.23)**	**pCR-615**
810	**β -galactosidase (EC 3.2.1.23)**	**pCR-810**
1383	**Gluconate dehydratase (EC 4.2.1.6)**	**pDr-1383**
596	**α-amylase (EC 3.2.1.1)**	**pCR-596**
985	**α-glucosidase (EC 3.2.1.20)**	**pCR-MalP**
421	**Glucose-1-dehydrogenase (EC 1.1.1.470)**	**pCR-GDH**
594	**Glycogen debranching enzyme**	**pCR-594**
595	**Glycosyl transferase (EC 2.4.1.-)**	**pDr-595**
960	1,4-**α**-glucan branching enzyme (EC 2.4.1.18)	pCR-960
844	**Thermopsin (EC 3.4.99.43)**	**pDr-844**
1054	**Thermopsin (EC 3.4.99.43)**	**pCR-1054**
506	Nitrilase (EC 3.5.5.1)	pCR-506
339	Carboxylesterase (EC 3.1.1.1)	pCR-339
1534	Glycerophosphoryl diester phosphodiesterase (EC 3.1.4.46)	pCR-1534
1065	Acetamidase (EC 3.5.1.4)	pCR-1065
1139	Amidohydrolase (EC 3.5.1.-)	pCR-1139

In general, the cloning of these ORFs was accomplished by PCR using *P. torridus* genomic DNA as a template and the primers listed in Table 4. The obtained PCR products were cloned either in pCR4_TOPO or pDrive vectors following the description of the manufacturer. For further subcloning of the genes into the expression vectors the fragments containing the cloned genes were derived by restriction using either the *Nde*I or *Nco*I sites introduced with the primers at the 5'-ends of the ORFs and one of the restriction sites present in the MCS of the vector downstream of the cloned gene. The acceptor vectors were subjected to restriction with the same enzymes, dephosphorilated and ligated with the gene containing fragments (section B.2.3)

However, there are several factors that can significantly influence the outcome of using this system.

It is well known that the codon usage of *E. coli* is highly biased. In particular, arginine AGA and AGG codons are extremely rare, which can severely affect the heterologous production of archaeal proteins, where these are the major codons for arginine. Also, synthesis of recombinant proteins in *E. coli* can often result in rapid degradation or aggregation of these proteins because of their inability to form native or correct tertiary structures. The first problem can be overcome by introducing genes into an expression system with elevated levels of tRNAs for these rare codons. Such an *E. coli* strain is a BL21 derivative – Rosetta, available from Novagen, and it was used in the expression experiments. For overcoming of the second problem different approaches can be undertaken. These are described below.

2.1. Refolding of solubilised inclusion bodies

In the form of inclusion bodies the target protein is insoluble, misfolded and inactive; thus, it is necessary to refold the protein to regain its activity. Active protein can be obtained from inclusion bodies by solubilising the aggregates in high concentrations of denaturants such as guanidine hydrochloride (GdnHCl; 4–7 M) or urea (5–10 M). After the dissolution, the misfolded polypeptide can refold into its native conformation when the denaturant is removed (Sambrook and Russel, 2001). It has to be noted, however, that this approach is highly empirical and laborious.

For these experiments, three *P. torridus* genes were used: ORFs 615 and 810, both annotated as β-glucosidase, and 596 coding for α-amylase. The genes were initially amplified by PCR using modified primers (see section B.2.4., Table 4) and genomic DNA as a template, cloned in pCR4-TOPO and further subcloned in the expression vector pET24c using *NdeI* and *NotI* restriction sites giving p24c-615, p24c-810 and p24c-596. The expression constructs were introduced in *E. coli* Rosetta and expression was induced by the addition of 0.1 mM IPTG to the growth medium. SDS-PAGE analysis of the different cell fractions (soluble and insoluble, see section B.1.4.) displayed abundant inclusion body formation in all three cases which is evident from the appearance of large amounts of recombinant protein in the insoluble fraction of crude extracts (Figure 12). Also, the soluble fractions displayed no enzyme activity in assays for β-galactosidase and α-amylase respectively.

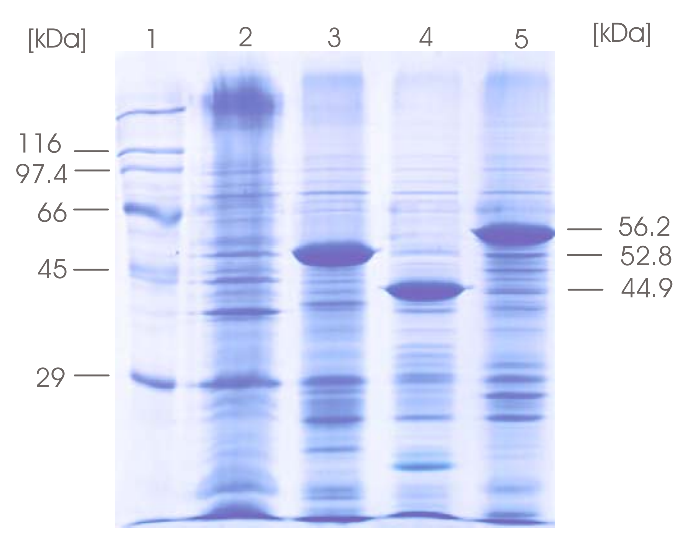


Fig. 12. SDS PAGE gel showing protein bands from inclusion bodies.(IF- insoluble fraction, R – Rosetta)

Lane 1: MWM

Lane 2: *E. coli*R pET24c IF

Lane 3: *E. coli*R p24-615 IF

Lane 4: *E. coli*R p24-810 IF

Lane 5: *E. coli*R p24-596 IF

In all the refolding methods applied, the inclusion bodies were purified (see section B.3.4.) and solubilized in 6 M guanidine HCl. An important indication of successful refolding in these cases would be the gaining of enzyme activity of the samples. Therefore enzyme activity measurements were performed with the protein samples. Another option for checking the state of the protein is to apply it to native SDS after denaturant removal. Only folded polypeptides would be able to enter the gel – high molecular weight aggregates should stay in the stacking gel.

2.1.1. Refolding by dialysis or rapid dilution

The critical factors that affect the outcome of these methods are the protein concentration and the composition of the refolding buffer (Vuillard *et al.*, 1998). Dialysis and rapid dilution were performed therefore with different protein concentrations of the inclusion body (IB) preparation (from 0.5 µg/ml to 10 µg/ml) and with different end concentrations of denaturant (0.01 – 0.2 M guanidine HCl) in the refolding buffer. Also, buffers with pH values ranging from 4 to 7 were tested (section B.3.4). In all three cases the dialysis resulted in precipitation of the protein. Moreover, the samples were negative for their corresponding activity and were not separable on a native PAGE (data not shown).

2.1.2. Refolding using the *Vectrase* kit

In these experiments, inclusion bodies obtained from the recombinant *E. coli* Rosetta clones (p24-615c and p24c-810) expressing the genes for the two probable β-galactosidases were used, designated as IB615 and IB810. The method screens for the refolding capacity of four different *Vectrase* detergents (1-4) and two *Vectrase* CDs (“stripping” agents, CD I and CD II). Thus eight different conditions are tested. Analysis by native PAGE showed that two of the conditions led to the formation of soluble forms of the proteins and these conditions were the same for both samples tested (Fig. 13).

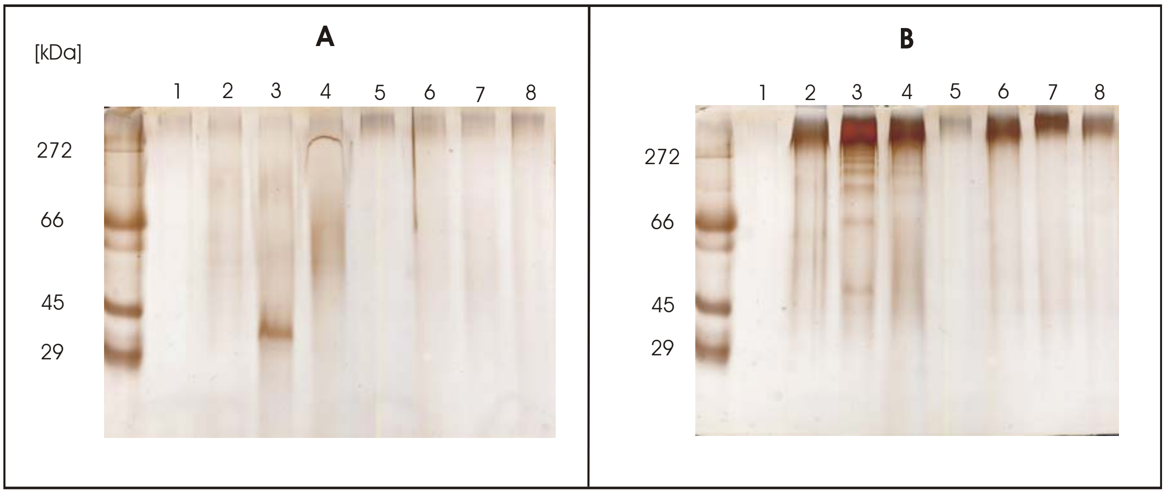


Fig. 13. Silver stained native PAGE of IB obtained from *E. coli*R p24-615 (A) and *E. coli*R p24-810 (B) after refolding with the *Vectrase* kit. In lanes from 1 to 8 are applied the samples from the eight different conditions, as described in the kit were applied to lanes 1 to 8: lanes 1÷4, CD I; lanes 5÷8, CD II.

The combination of “stripping” agents and detergents that led to the formation of soluble proteins was *Vectrase* CD I and detergent 3 for both “inclusion bodies” proteins tested. While in the case of IB615 the soluble polypeptide had the form of a single band, IB810 gave multiple bands on the native gel which is probably due to the formation multimers with different number of monomers. However, both proteins were inactive when tested for enzyme activity.

2.1.3. Refolding using size exclusion chromatography (SEC)

The use of a gel filtration material as a medium for refolding of proteins has been described (for a review of this approach see Chaudhuri, 1994) and several examples of successful refolding have been published (Batas *et al.,* 1999, Harrowing *et al.*, 2003). The gel medium allows refolding to occur even at high protein concentrations by preventing inter-molecular interactions and at the same time serves to separate the refolded species from the misfolded or aggregated ones and to remove the denaturant.

Figure 14 shows the behaviour of IB615 and IB810 during such a size exclusion refolding experiment. Under these conditions, high molecular weight aggregates elute at V_0, which for this column is 39.7 ml.

For these experiments the samples used were the same as in the previous method described, i.e. IB615 and IB810. 5 ml of each IB preparation (see section B.3.4.), containing 6 M guanidine HCl, were applied on a Superdex 200 16/60 column equilibrated with 50 mM phosphate buffer pH 6.5. A standard sample consisting of 3 mg each of cytochrome C, egg albumin and katalase was run under the same conditions. The total amount of protein applied was 4.5 mg for IB615 and 10 mg for IB810. The final denaturant concentration after elution was calculated to be 0.2 M.

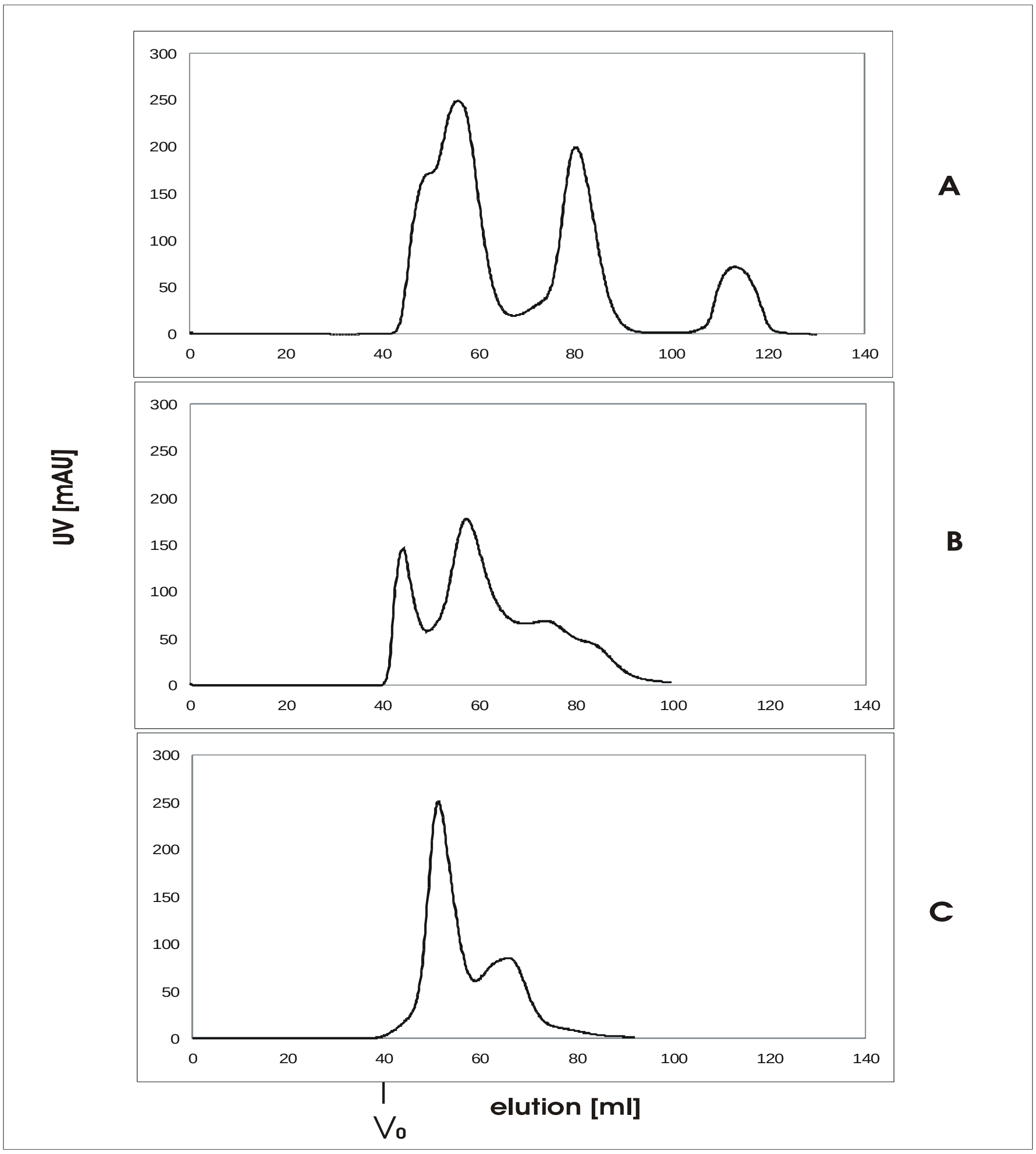


Fig. 14. Size exclusion chromatography refolding of IB615 (B) and IB 810 (C). The molecular weight standard (A) consists of 3 mg each of cytochrome C (12.5 kDa), egg albumin (45 kDa) ,katalase (240 kDa).

After 1.2 column volumes the fractions were analyzed by native PAGE (not shown) and checked for enzyme activity. The analyzed fractions showed no activity with the substrates para-nitrophenyl β D-galactoside, para-nitrophenyl α D-glucoside or para-nitrophenyl β D-glucoside . Although the peaks of the IB samples indicated soluble form of the protein, *i.e.* eluting after the V_0 of the column, the lack of enzyme activity suggested that these forms were not the biologically native ones.

2.2 Use of a weak promoter

A possible reason for the abundant inclusion body formation that was observed is the use of an expression system with a very strong promoter, i.e. T7. To test this possibility an alternative expression vector was used, pBAD/*Myc*-His, based on the *E. coli* arabinose promoter (Invitrogen). This vector system allows a more precise, dose-dependent regulation of the expression level by varying the concentration of the inducer (L-arabinose) (Schleif, 2000). A general strategy for subcloning into this vector was the use of the *Nco*I site introduced during the amplification of the corresponding gene by modified primers. The ORFs which were subcloned in pBAD/*Myc*-His and their assigned function, as well as the restriction enzymes used for this purpose, the number of amino acids of the encoded protein and the names of the resulting constructs are shown in Table 11.

Table 11. ORFs cloned in the expression vector pBAD/*Myc*-His.

ORF No	Assigned function	RE used for subcloning in pBAD/*Myc*-His	Number AA	Construct name
421	Glucose-1-dehydrogenase	*Nco*I	359	pBAD-GDH
594	Glycogen debranching enzyme	*Nco*I – *Kpn*I	577	pBAD-594
595	1,4-α-glucan branching enzyme	*Nco*I - *Kpn*I	376	pBAD-595
844	Thermopsin	*Nco*I	449	pBAD-844_1
1054	Thermopsin	*Nco*I - *Xho*I	722	pBAD-1054
506	Nitrilase	*Nco*I - *Eco*RI	240	pBAD-506

Two *E. coli* strains were transformed with each of these expression constructs, TOP10 (T10) and Rosetta. *E. coli* T10 is capable of transporting L-arabinose (*ara*EFGH$^+$) but on the other hand carries several mutations which prevent metabolising it (*ara*BACD$^-$) and in this way constant levels of the inducer in the cell can be maintained. In the case of *E. coli* Rosetta which is *ara*BACD$^+$ induction of gene expression was achieved with a maximal concentration of inducer (0.2 % w/v), while when expressing from the T10 strains varying amounts of L-arabinose were applied ranging from 0.00002 to 0.2 % w/v in order to optimise the expression level. After growing the transformed cells for 6 h after induction, crude cellular fractions were analyzed for the presence of recombinant protein by SDS-PAGE and enzyme activity assays. Only *E. coli* Rosetta pBAD-GDH was found to express the expected protein and its presence was confirmed also by measuring glucose dehydrogenase activity (10 $U \cdot mg^{-1}$). The recombinantly expressed *P. torridus* glucose-1-dehydrogenase (GdhA) was subjected to purification and biochemical characterisation (see section C.3).

2.3. Expression of fusion proteins

Another approach that could lead to the expression of soluble recombinant proteins in *E. coli* is to use fusion partners – to clone the gene of interest in frame with a *E. coli* gene whose product is known to be highly soluble. After obtaining the fusion product the two proteins are separated by cleavage with a site specific protease whose recognition sequence is encoded in the vector 5’ to the polylinker region. Two such fusion systems were used: pET43.1 (Novagen) where the fusion partner is the 495 aa NusA protein, and pMAL-c2x (NEB) which leads to fusion with the 462 aa maltose-binding protein (MBP) from *E. coli*.

2.3.1. Maltose binding protein tag – pMAL-c2x

Two *P. torridus* genes were cloned in this vector: ORF 1383 with an assigned function gluconate dehydratase and ORF 810 annotated as β-galactosidase. Cloning of ORF 1383 was accomplished by restricting the vector pDr-1383 (pDrive with PCR amplified ORF 1383 cloned into it) with *BamH*I and *Xba*I and cloning the fragment containing the gluconate dehydratase gene in pMAL-c2x cut with the same enzymes, giving pM1383. For cloning of ORF 810 a PCR amplification of the gene was done using primers 810for_nde and ABI_for and the plasmid pCR4-810 (described in section C.2.1.) as a template. The blunt-ended PCR product was cut with *Hind*III (giving a DNA fragment with a blunt 3’ end and *Hind* III compatible 5’ overhang) and ligated with the pMAL-c2x vector restricted with *Xmn*I and *Hind*III resulting in pM810.

The constructs were introduced in *E. coli* Rosetta and expression from the P_{TAC} promoter was induced with 0.1 mM IPTG after the OD of the cell culture had reached

0.5. The soluble fractions were analysed for the presence of fusion protein by SDS PAGE. Both strains with the fusion constructs yielded soluble proteins of the expected molecular mass (see Fig. 15 and 16). The samples were further partially purified with an affinity matrix amylose using the property of MBP to bind maltose (see section B.3.5.3).

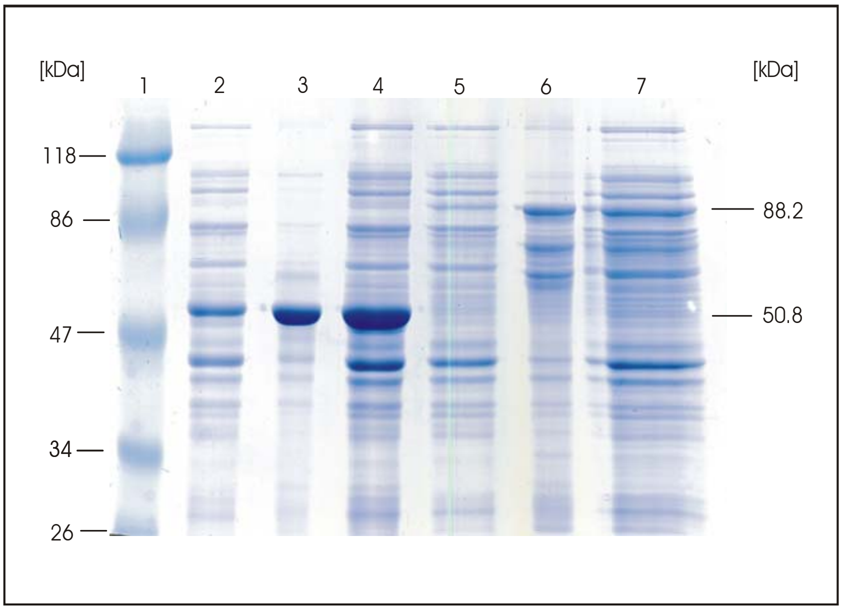


Fig. 15. SDS-PAGE of the soluble fractions of *E. coli* Rosetta pMALc2x and pM1383.

Lane1: MWM
Lane2: pMALc2x W
Lane3: pMALc2x Amy
Lane4: pMALc2x S
Lane5: pM1383 W
Lane6: pM1383 Amy
Lane7: pM1383 S

W – wash fraction after purification with amylose.

Amy – amylose binding fraction.

S – soluble fraction.

In the pMALc2x system a specific protease – FactorXa (NEB) is used to cleave the fusion product in order to separate the two proteins. For the case of pM810 the sample was further subjected to FactorXa cleavage and the products were analyzed by SDS-PAGE. The molecular mass of the proteins obtained after cleavage was in agreement with the predicted one (Fig. 16B).

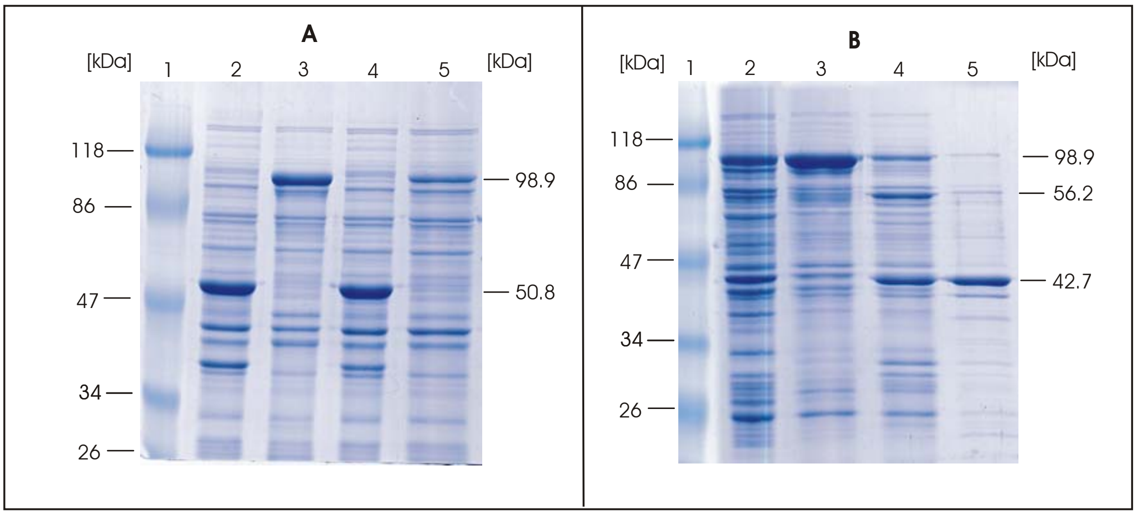


Fig. 16. (A) SDS-PAGE of soluble and crude extract fractions of *E. coli* Rosetta pMALc2x and pM810. (B) SDS-PAGE of *E. coli* Rosetta pM810 protein fractions before and after FactorXa cleavage. S-soluble fraction; CE-crude cellular extract

Lane1: MWM
Lane2: pMALc2x S
Lane3: pM810 S
Lane4: pMALc2x CE
Lane5: pM810 CE

Lane1: MWM
Lane2: pM810 S
Lane3: pM810 amylose-binding fraction
Lane4: same as lane 3, after FactorXa treatment
Lane5: same as lane4 heat treated 60°C 10 min

When tested for the corresponding enzyme activity, both proteins were negative before as well as after FactorXa cleavage. As in the refolding experiments (section C.2.1) the soluble protein species observed were probably not the native ones.

2.3.2. NusA tag – pET43.1

P. torridus ORF 810 was cloned in frame with the fusion tag NusA of the pET43.1 vector by restricting the pCR4-810 vector with *Nde*I, filling in of the protruding termini with Klenow enzyme followed by *Hind*III digestion. This procedure gave a DNA fragment with a blunt 3' end and a *Hind*III-compatible 5' overhang which was ligated with a *Psh*AI-*Hind*III-restricted expression vector. The construct obtained was termed pN-810. Figure 17 shows SDS-PAGE analysis of the soluble and crude extract fractions of *E. coli* Rosetta transformed with pET43.1 as a control and pN-810 after induction with 0.1 mM IPTG.

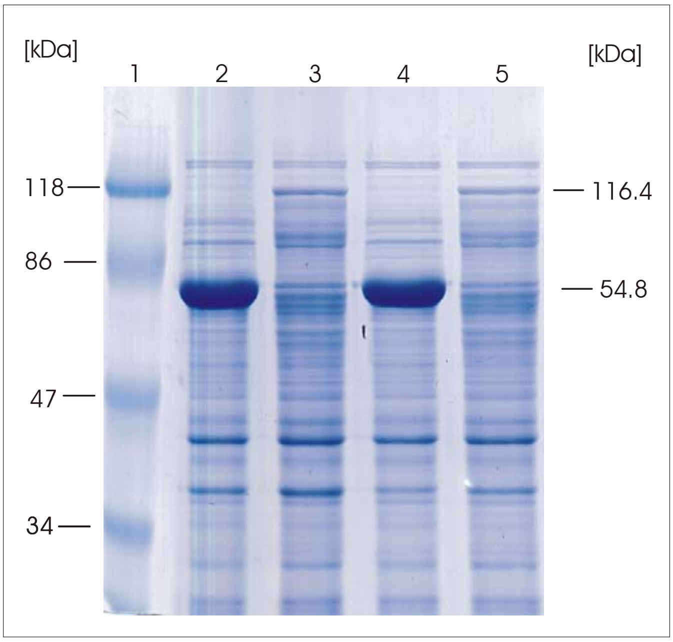


Fig. 17. SDS-PAGE of the fractions of *E. coli* Rosetta pET43.1 and pN-810

Lane1: MWM
Lane2: pET43.1 S
Lane3: pN-810 S
Lane4: pET43.1 CE
Lane5: pN-810 CE

S – soluble fraction
CE – crude cellular extract

The fusion protein had a size corresponding with the predicted one but also did not display activity with the substrates that were routinely used when screening for β-galactosidase activity.

2.4. Other expression systems – the SSV1 virus and *S. solfataricus*

One of the few archaeal genetic systems that have been developed is based on the SSV1 virus (*Sulfolobus shibatae* virus 1) (Palm *et al.*, 1991). Derivatives of this virus have been constructed that can be propagated in *E. coli* at high copy number and spread efficiently through infected cultures of *S. solfataricus* both specifically integrated in the host chromosome and as circular episome (Stedman *et al.*, 1999). In this work, pMJ03 was used (Jonuscheit *et al.*, 2003), a SSV1 derivative that carries the *lac*S gene from *S. solfataricus* under the control of the heat-inducible *tf*55 promoter. *P. torridus* ORF 810 was cloned in pMJ03 by using a PCR fusion strategy which is schematically shown in Figure 18.

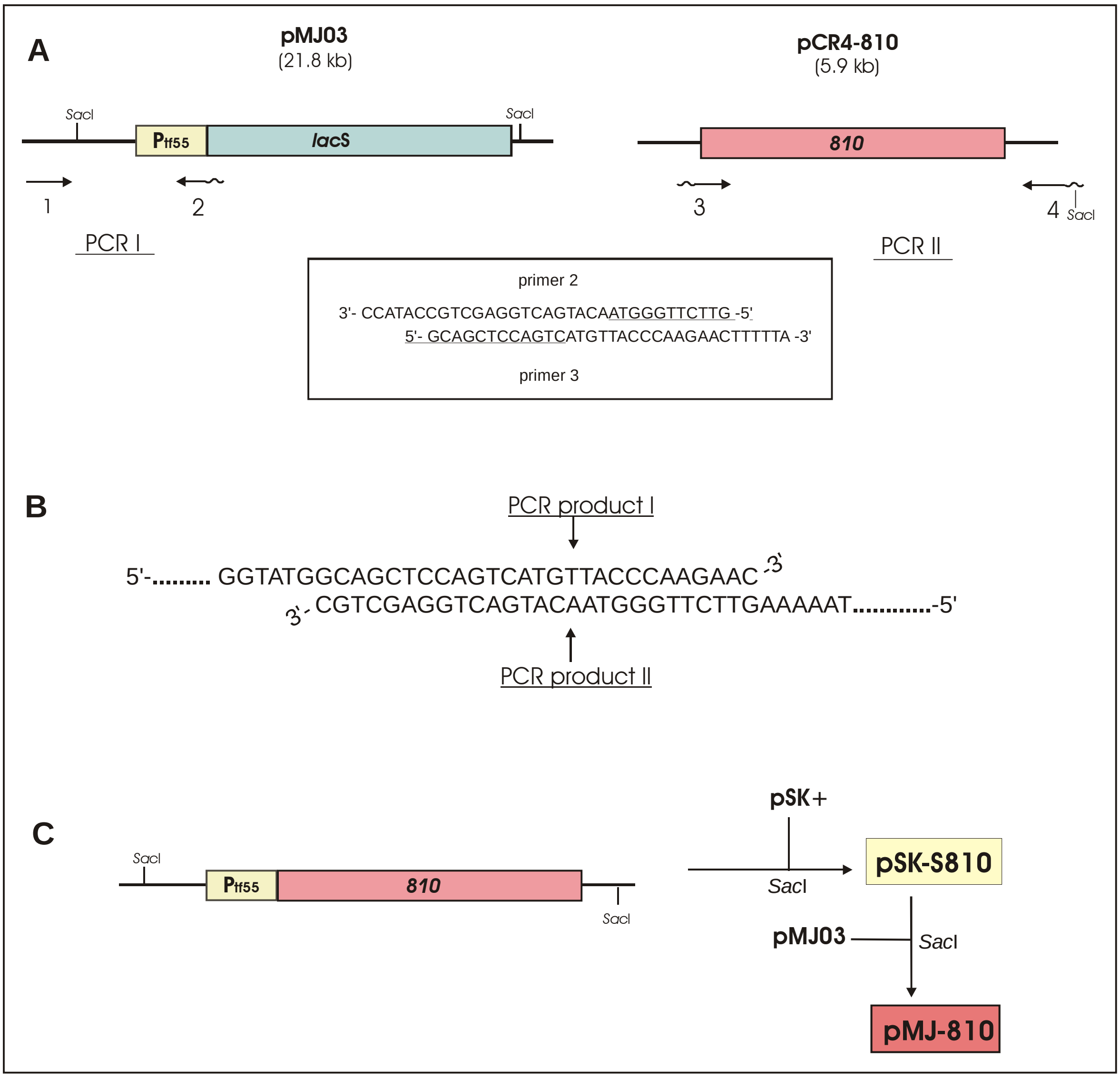

A
pMJ03
(21.8 kb)
pCR4-810
(5.9 kb)
SacI
SacI
Ptf55
lacS
810
1
2
3
4
SacI
PCR I
PCR II
primer 2
3'- CCATACCGTCGAGGTCAGTACAATGGGTTCTTG -5'
5'- GCAGCTCCAGTCATGTTACCCAAGAACTTTTTA -3'
primer 3
B
PCR product I
5'-......... GGTATGGCAGCTCCAGTCATGTTACCCAAGAAC -3'
3'- CGTCGAGGTCAGTACAATGGGTTCTTGAAAAAT............-5'
PCR product II
C
SacI
Ptf55
810
SacI
pSK+
SacI
pSK-S810
pMJ03
SacI
pMJ-810

Figure 18. Flow-chart of the cloning strategy used for placing *P. torridus* ORF 810 in the viral vector pMJ03 under the control of the *tf55* promoter.

A. In the first step, two PCR products were generated (I and II) using primers 1-2 and 3-4 respectively. Primers 2 and 3 have 5' overhangs which are not complementary to the template (underlined) and overlap like it is shown in the box.

B. In the second step, PCR I and II were used as a template for a second PCR round initially without primers (5 cycles). After the first denaturing step the two molecules anneal like it is shown due to the complementary sequence introduced by the primers in the first stage. This leads to the formation of a fusion product which was amplified by adding primers 1 and 4 to the reaction and performing additional 25 cycles.

C. The obtained fusion PCR product was digested with *Sac*I (the 5' *Sac*I site is introduced with primer 4) and cloned in pSK+ restricted with the same enzyme. The *tf55-810* construct was further moved to pMJ03 with the use of the same enzyme resulting in pMJ-810.

The primers used were as follows (see section B.2.4): 1- pyrEF.F, 2- tf55-1070.R, 3- S1070.F, 4- ABI.for-sacI

S. solfataricus PH1-16 was transformed with the obtained plasmid (pMJ-810) as described in section B.2.5.2 (see also Schleper *et al.*, 1992). This strain was used because its phenotype (Lac$^-$) would permit selection of positive transformants based on the presence of β-galactosidase activity (Lac$^+$). Successful transformation was confirmed by PCR (not shown). However, no β-galactosidase activity could be detected after 4 days of incubation of the primary transformation mixture at 75°C. Also, on SDS-PAGE gels there was no detectable difference between the transformed cells and the negative control (not shown).

2.5. Expression in *S. cerevisiae*

For expression attempts in *S. cerevisiae* the vector pYes2 NT-A was used. ORF 810 was amplified from *P. torridus* genomic DNA using the primers Y810.F_BH and Y810.R_BH and the PCR product was cloned in pCR4_TOPO resulting in pCR_Y810. Subcloning into pYes2 NT-A was accomplished with the use of the *Bam*HI restriction sites of pCR_Y810 introduced with the primers and the *Bam*HI restriction site present in the MCS of pYes2 NT-A. This resulted in a fusion between the His tag present in the acceptor vector and the N-terminus of ORF 810.

However, after introduction of the expression construct in the *S. cerevisiae* strain INVSc1 and induction with galactose, no detectable level of β-galactosidase activity could be measured.

3. Purification and biochemical characterisation of glucose-1-dehydrogenase and α-glucosidase from *P. torridus*

3.1. *P. torridus* α-glucosidase (MalP)

3.1.1. Analysis of the amino acid sequence of MalP

P. torridus ORF 985 (*malP*) was annotated as α-glucosidase (EC 3.2.1.20) based on its amino acid sequence similarity (38.6 % identity) with the characterised homolog from *S. solfataricus*. The protein consisted of 645 aa and could be classified as a member of glycoside hydrolase family 31 based on the presence of the conserved active site pattern [GF]-[LIVMF]-W-x-D-M-[NSA]-E characteristic for this family (CAZy database, Henrissat, 1991). At present this family contains numerous (184) enzymes belonging to organisms from all major branches of the phylogenetic tree, among them 6 of archaeal origin, having several known activities: α-glucosidase (EC 3.2.1.20), α-galactosidase (EC 3.2.1.22); glucoamylase (EC 3.2.1.3), sucrase-isomaltase (EC 3.2.1.48) (EC 3.2.1.10); α-xylosidase (EC 3.2.1.-) and α-glucan lyase (EC 4.2.2.13).

Analysis of the genome sequence surrounding the *mal*P gene revealed the presence of a gene coding for an α-mannosidase located upstream of *mal*P. Interestingly, such a functional cluster was found only in *P. torridus* and in none of the organisms that possess homologs of one of the both ORFs (Figure 19).

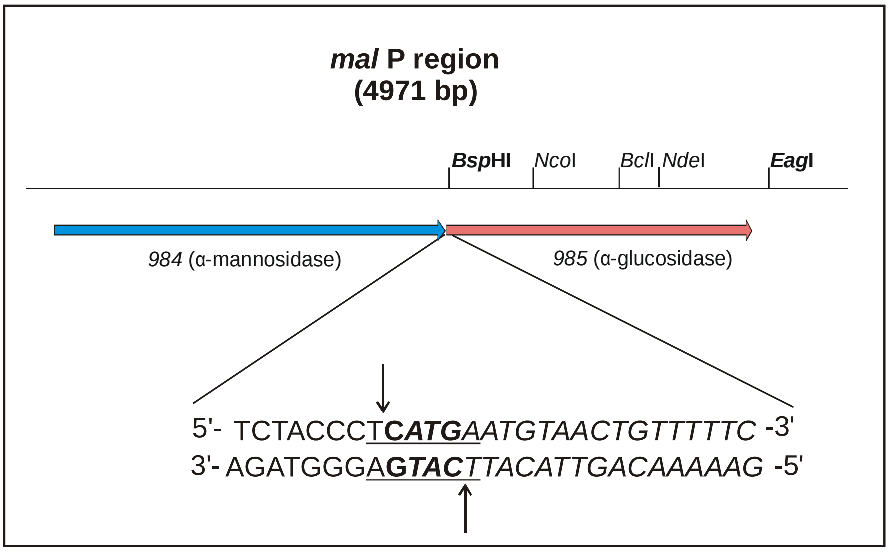


Fig. 19. The *P. torridus* genome region containing the *malP* gene and restriction enzymes used for its cloning into pET24d. The enzymes used for cloning of the gene are in bold. The MalP coding sequence is shown in italic, the *Bsp*HI recognition site is underlined and the *Nco*I-compatible overhangs generated after *Bsp*HI cleavage are in bold.

3.1.2. Cloning and expression of the *P. torridus mal*P ORF and purification of recombinant MalP

Suitable restriction enzymes were found in the native sequence which were later used for subcloning of *mal*P into the expression vector pET24d (Fig. 17). Cloning of the *P. torridus* ORF 985 was accomplished by amplifying the region surrounding the *malP* gene with the primers 985reg_for and 985reg_rev using genomic DNA as the template. The 2996 bp product contained 500 bp DNA sequence upstream of the gene's start codon (ATG) and 500 bp sequence downstream of its stop codon (amber) and was cloned in pCR4-TOPO using the Topo cloning kit (Invitrogen). The vector obtained (pCR-985) was cut with *Bsp*HI (*Nco*I-compatible) and *Eag*I (*Not*I-compatible) and the fragment containing the *985* gene was ligated with pET24d digested with *Nco*I and *Not*I. The resulting construct was termed p24-malP and was used to transform *E. coli* Rosetta.

Analysis of *E. coli* Rosetta cells transformed with p24-malP showed the presence of α-glucosidase activity (107 $mU \cdot mg^{-1}$) when tested with para-nitrophenyl α-D-glucoside which did not decrease even after heating the cell fractions at 70°C for 15 min. The expression of a soluble, heat stable α-glucosidase having the predicted molecular mass was also confirmed by SDS-PAGE (see figure 20).

The MalP protein was purified from 4 l of *E. coli* Rosetta p24-malP culture by a three stage process which makes use of the observed thermostability of the enzyme.

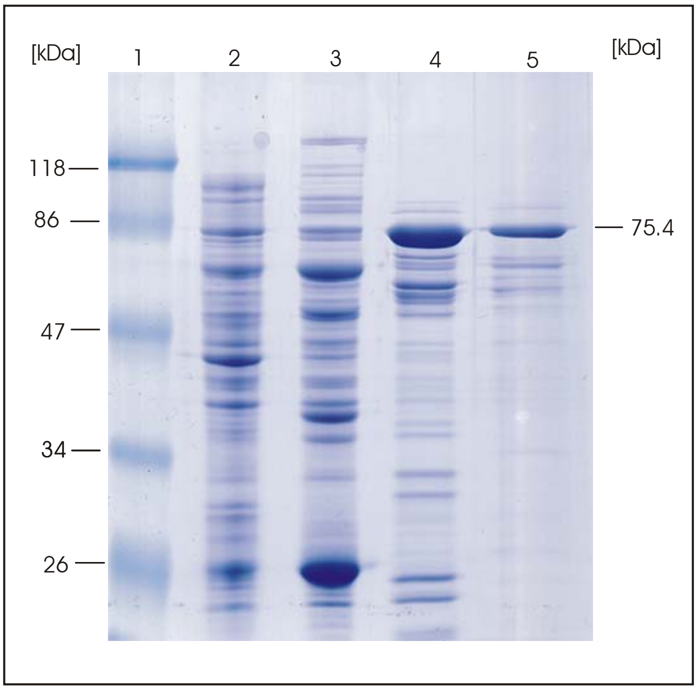


Fig. 20. SDS PAGE showing the stages in the purification of MalP.

Lane1: MWM
Lane2: *E. coli* p24-malP CE
Lane3: *E. coli* p24-malP HT
Lane4: Pooled fractions after AIEX.
Lane5: Pooled fractions after SEC

CE – crude extract
HT – heat treated fraction
AIEX – anion exchange chromatography (Source Q30)
SEC – size exclusion chromatography (Superdex 200)

As a first step in the purification, the cellular extract was subjected to heat treatment for 15 min at 70°C, the precipitated proteins collected by centrifugation (13,000 x g, 20 min, 4°C) and after filtering (0.22 µm filter) the supernatant was loaded on a Source Q30 anion exchange column. After determining the enzyme activity of the collected fractions, the three most active ones were pooled (9 ml), concentrated to 1ml and applied on a Superdex200 gel filtration column. A detailed description of the anion exchange and size exclusion chromatography steps in the purification of MalP can be found in section B.3.5.2. Table 12 and Figure 20 show the purification steps of MalP.

Table 12. Purification of MalP. Activity was measured with pNPαD-glucoside as a substrate by incubation for 10 min at 85°C (see section B.3.6.2). Specific activity is expressed as µmol para-nitrophenol produced/min/mg of protein

Step	**Total protein (mg)**	**Total activity (U)**	**Specific activity ($U \cdot mg^{-1}$)**	**Yield (%)**	**Purification (fold)**
Cell-free extract	1721	184	0.11	100	
Heat treated	386	115	0.29	62.5	2.6
Source Q30	41	93	2.2	50.5	20
Superdex 200	1.3	68	54.2	36.9	492

3.1.3. Temperature, pH optimum and reaction products of MalP

Using a 10 min assay, the purified recombinant α-glucosidase was found to be most active at 87°C when tested with para-nitrophenyl α-D-glucoside in 50 mM acetate buffer at pH 5.0 (Fig. 21B). Surprisingly, this value is 27°C above the optimum growth temperature of *P. torridus* and more than 20°C above its maximum growth temperature. The enzyme displayed significant activity over a broad pH range (Fig. 21A). The pH optimum of MalP activity corresponds well to the intracellular pH of the *P. torridus* cells (pH 4.6) Thin layer chromatography analyses (TLC) of the reaction products showed that exclusively glucose was formed when malto-oligosaccharides with different chain lengths were used as substrates (Fig 22)..

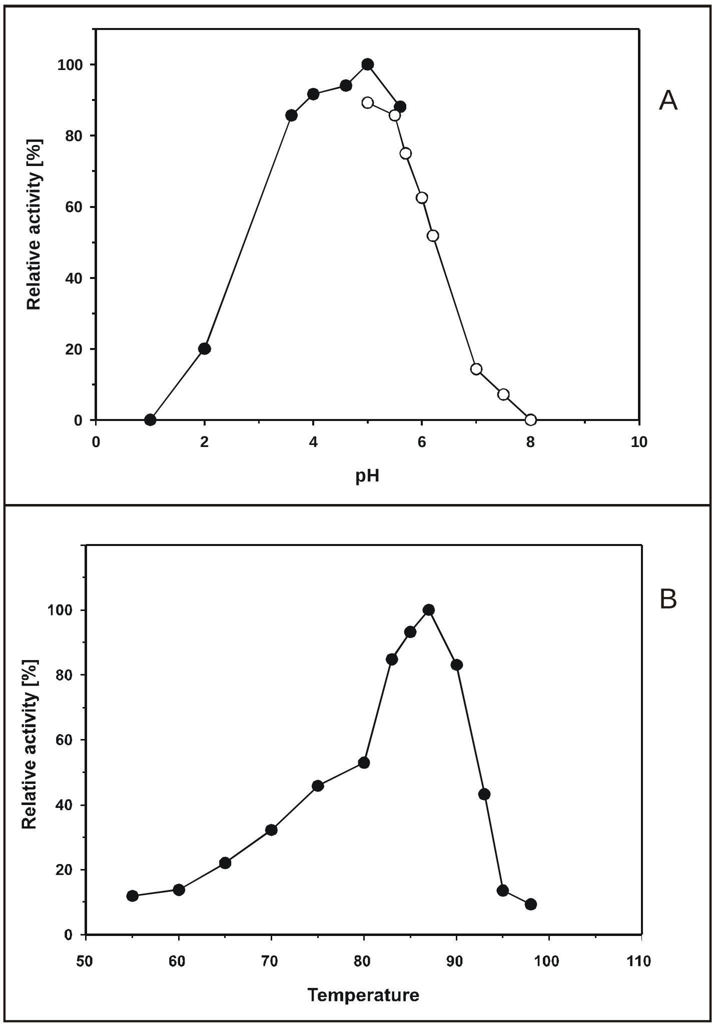


Figure 21. pH and temperature dependence of recombinant MalP activity. Relative activity is expressed as [%] of the maximum.

A. pH optimum was determined at 87°C in 50 mM acetate buffer (black circles) and McIlvaine buffer (white circles).

B. The temperature dependence of activity was determined with a 10 min assay in 50 mM acetate buffer pH 5.0 and 10 mM pNP α-D-glucoside.

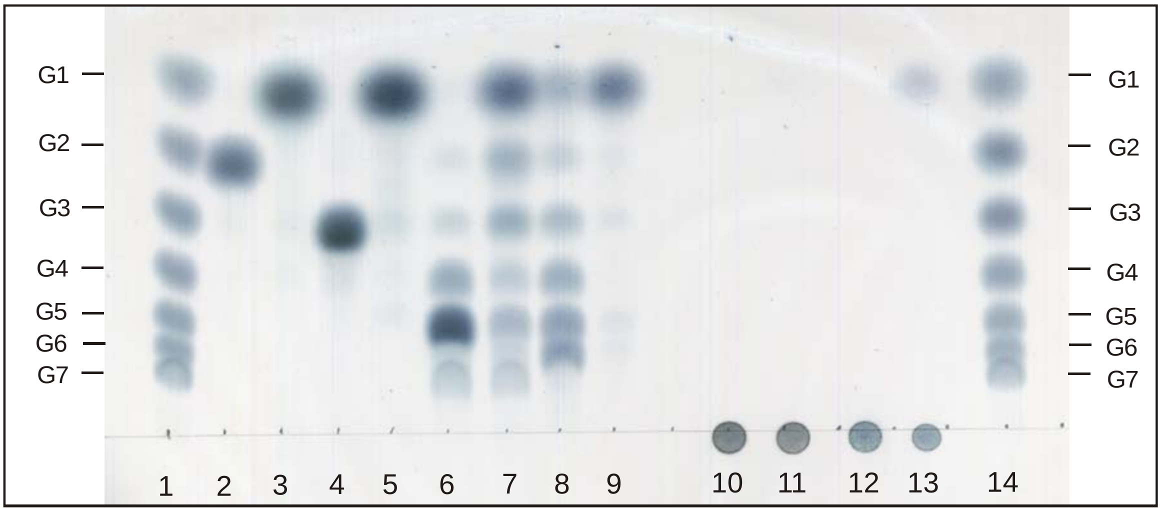


Fig. 22. TLC analysis of the reaction products of MalP. The reactions were carried out in 100 µl with 3.6 µg purified MalP in 50 mM acetate buffer pH 4.5 containing 10 mM substrate or 0.3 % dextrin and starch. G1-glucose, G2-maltose, G3-maltotriose and s.o.

Lanes 1 and 14	**maltooligosaccharide standard**	**Lanes 8 and 9**	**maltohexose [+] and [-] enzyme**
Lanes 2 and 3	**maltose [+] and [-] enzyme**	**Lanes 10 and 11**	**dextran [+] and [-] enzyme**
Lanes 4 and 5	**maltotriose [+] and [-] enzyme**	**Lanes 12 and 13**	**starch [+] and [-] enzyme**
Lanes 6 and 7	**maltopentose [+] and [-] enzyme**		

A further characterisation of MalP was not carried out in the time frame of the current work.

3.2. *P. torridus* glucose-1-dehydrogenase (GdhA)

Metabolic pathway reconstruction based on the genome data suggested the presence of a non-phosphorylated variant of the Entner-Doudoroff pathway (section C.1.3.9, Fig. 9). For its first enzyme, glucose dehydrogenase (EC 1.1.1.47), three ORFs were identified in the annotated *P. torridus* genome, each coding for different proteins with similarity to glucose dehydrogenases belonging to the medium-chain alcohol dehydrogenase family (data not shown). ORF 421 showed a high degree of amino acid sequence similarity to the glucose dehydrogenase from *T. acidophilum* (59.3 % identity) and was selected for cloning and heterologous expression.

3.2.1 Analysis of the amino acid sequence of GdhA

This ORF codes for a protein of 359 amino acids (M_r 40,462) which corresponds in size to the purified enzyme as determined by SDS-PAGE (Fig. 24). Based on amino acid sequence similarity, *P. torrridus* glucose dehydrogenase could be assigned as a member of the medium–chain alcohol/polyol dehydrogenase/reductase branch of the superfamily of pyridine-nucleotide-dependent alcohol/polyol/sugar dehydrogenases (Edwards *et al.*, 1996). Members of this group are characterized by conserved structural and catalytic zinc-binding and nucleotide-binding sites. The crystal structure of the glucose dehydrogenase from *T. acidophilum* has been reported and the residues involved in zinc binding have been identified (John *et al.*, 1994). In the eukaryotic structural homologs of Gdh from *T. acidophilum* the structural zinc is coordinated by four cysteine residues that are highly conserved throughout the structural zinc-containing alcohol dehydrogenases, while the enzymes from *T. acidophilum* as well as *P. torridus* carry only three cysteine residues in this region. The fourth ligand has been established in *T. acidophilum* to be Asp 115, and the amino acid sequence alignment shows that *P. torridus* GdhA also has Asp at this position (Fig 23).

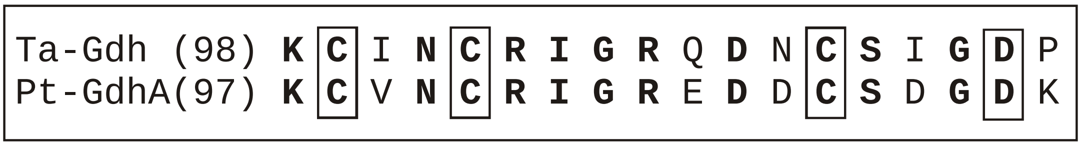

Ta-Gdh (98) K C I N C R I G R Q D N C S I G D P
Pt-GdhA(97) K C V N C R I G R E D D C S D G D K

Fig. 23. Amino acid sequence alignment of the regions involved in structural zinc binding of *P. torridus* GdhA and *T. acidophilum* Gdh. The numbers in brackets show the amino acid number that precedes the depicted ones. Identical amino acid residues are in bold, and the ones involved in zinc coordination (John *et al.*, 1994) are boxed.

In addition, the residues reported to be involved in Zn^{2+} coordination in the catalytic zinc binding region of the *T. acidophilum* glucose dehydrogenase (John *et al*., 1994) were also found in the primary structure of the *P. torridus* enzyme (not shown). The GXGXXG/A fingerprint motif, characteristic for pyridine nucleotide–binding proteins is also present, together with asparagine and histidine residues at positions 213 and 215 (*P. torridus* glucose dehydrogenase numbering), which are reported to explain the dual cofactor specificity of the enzyme from *T. acidophilum*.

3.2.2. Cloning and expression of the *P. torridus gdh*A ORF and purification of recombinant GdhA

Primers were constructed using the data of the *P. torridus* genome sequence (421for and 421rev) and gene amplification was accomplished by PCR with genomic DNA as template. The product was cloned in pCR4-TOPO and subsequently in pBAD/*Myc*His for expression giving pBAD-GDH (see section C.2.2). Presumably due to the presence of rare codons in the coding sequence of *gdhA* (most notably the Arg codon AGG with 3.3%), initial expression experiments in the *E.coli* strain TOP 10 carrying pBAD-GDH showed no detectable level of GdhA expression (data not shown). When the construct was introduced in *E. coli* Rosetta a high level of glucose dehydrogenase activity could be detected after induction with 0.2 % D-arabinose. The activity observed in the recombinant cells ($10\ U{\cdot}mg^{-1}$) was 700 fold higher than that in negative controls ($0.014\ U{\cdot}mg^{-1}$). Also, a higher level of expression was observed when the expressing *E. coli* cells were grown at 30°C compared to 37°C (not shown).

As an alternative, another expression vector was constructed – p24-GDH, which was obtained by subcloning the *gdhA* gene under control of the regulated T7 promoter of the vector pET24d. However, expression from this construct in *E .coli* Rosetta resulted in massive inclusion body formation.

The recombinant *P. torridus* glucose dehydrogenase was purified from *E.coli* Rosetta transformed with pBAD-GDH in a three stage process, which is summarized in Table 13. The thermostability of the enzyme permitted the use of heat treatment as a first step in the purification. By subsequent anion exchange and size exclusion chromatography the enzyme was purified to electrophoretic homogeneity. The isolated enzyme had a specific activity of 252 $U \cdot mg^{-1}$ and gave a single band on SDS/PAGE with a M_r corresponding to the size predicted from sequence analysis (Fig. 24). Gel filtration of the purified enzyme indicated a tetrameric structure (*Mr* approximately 160,000) which was not affected by the absence of NAD^+ or $NADP^+$ (not shown).

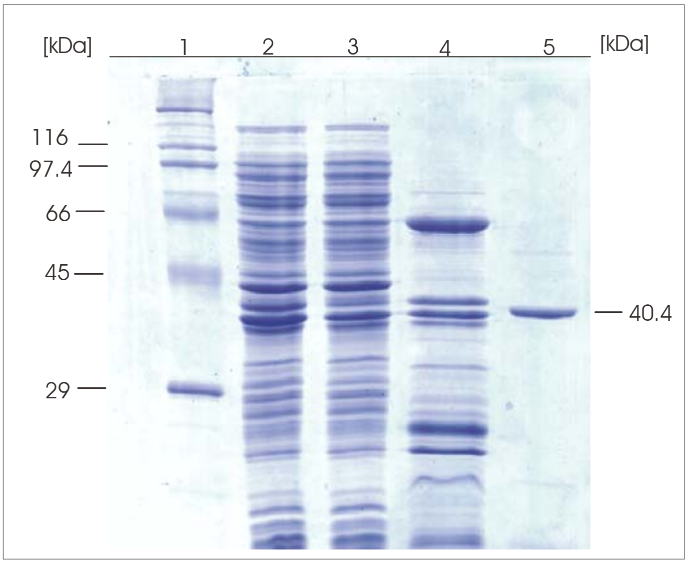


Fig. 24. SDS PAGE showing the steps in the purification of GdhA from *E. coli* Rosetta pBAD-GDH.
Lane1: MWM.
Lane2: pBAD-GDH CE
Lane3: pBAD-GDH S
Lane4: pBAD-GDH HT
Lane5: pBAD-GDH AIEX

CE – crude extract
S – soluble fraction
HT – heat treated fraction
AIEX–anion exchange chromatography

Table 13. The steps in the purification of recombinant GdhA. Activity was measured in 50 mM phosphate buffer pH 6.5 in the presence of 50mM D-glucose and 2 mM $NADP^+$ at 55°C for 10 min. Specific activity is defined as µmol NADPH produced/min/mg protein.

Step	Total protein (mg)	Total activity (U)	Specific activity ($U \cdot mg^{-1}$)	Yield (%)	Purification (fold)
Cell-free extract	72	720	10	100	1
Heat treated	26	689	26.5	96	2.6
Source Q	1.56	301	193	42	19
Superdex 200	0.6	151	252	20	25

3.2.3. Characterisation of recombinant *P. torridus* GdhA

3.2.3.1. Substrate specificity and effect of metabolites on the enzyme activity

When glucose was used as a substrate the purified enzyme had a specific activity of 252 U/mg with $NADP^+$ as a cofactor and 12.9 U/mg with NAD^+. A broad range of aldose sugars were tested as potential substrates for *P. torrridus* glucose dehydrogenase. The enzyme was significantly active only with D-galactose, reaching 74 % of the activity with D-glucose when $NADP^+$ was used as a cosubstrate. None of the C2 and C3 epimers of D-glucose or derivatives (D-mannose, D-allose, D-glucosamine, 2-deoxy-D-glucose, glucose-6-phosphate) and none of the aldopentoses (D-xylose, L-arabinose, D-ribose) tested showed activity above 2 % both with $NADP^+$ and NAD^+ as cosubstrates (Table 14). Interestingly, the enzyme showed higher activity with D-galactose when NAD^+ was used as a cosubstrate.

Table 14. Substrate specificity of recombinant GdhA.
The activity is expressed as [%] of the activity of D-glucose with $NADP^+$ and NAD^+ respectively. The standart assay (55°C, 10 min, phosphate buffer pH 6.5) contained 40 mM of each substrate and 2 mM $NADP^+$ (5mM NAD^+).

	Cosubstrate	
Substrate	**$NADP^+$**	**NAD^+**
D-Glucose	100	100
D-Mannose	3.1	3
D-Galactose	74.7	112
D-Fructose	1.0	5
D-Xylose	2.4	1
D-Fucose	0	0
D-Glucosamine	0.5	2
D-Ribose	0	1
D-Arabinose	0.4	1
Glucose-6-phosphate	3.9	0
D-Lactose	0.4	1
D-Allose	0	0
6-deoxy- D-glucose	0	0
2-deoxy- D-glucose	0	0

Additionally, the influence of adenine nucleotides, inorganic phosphate and pyrophosphate, and downstream products of the Entner–Doudoroff pathway on enzymatic activity was tested in order to investigate whether GdhA was regulated by metabolites or the energy status of the cell. The enzyme was found to be affected only by ATP, which led to a decrease in activity of 30 % at 5 mM concentration in the reaction buffer. Pyruvate, phospho-enol-pyruvate, 3-phospho-glycerate, 2-phospho-glycerate, as well as Pi and PPi did not significantly affect the activity.

3.2.3.2. Kinetic properties of GdhA

The recombinant *P. torridus* glucose dehydrogenase was active with glucose and galactose and both $NADP^+$ and NAD^+ as cosubstrates, displaying approximately 20-fold higher activity with $NADP^+$. Kinetic analysis, accomplished by the direct linear plot method with glucose and $NADP^+$ as substrates, resulted in K_m values of 10 mM (± 1) for glucose and 1.12 mM (± 0.2) for $NADP^+$. When NAD^+ was used as a cosubstrate, the affinity of the enzyme for glucose was significantly lower (K_m = 88 mM ± 5). Exact determination of the Km value for NAD^+ using D-glucose as a substrate was not possible obviously because of the weak affinity of the enzyme towards NAD^+. Even cosubstrate concentrations above 30 mM could not lead to enzyme saturation (Fig. 25).

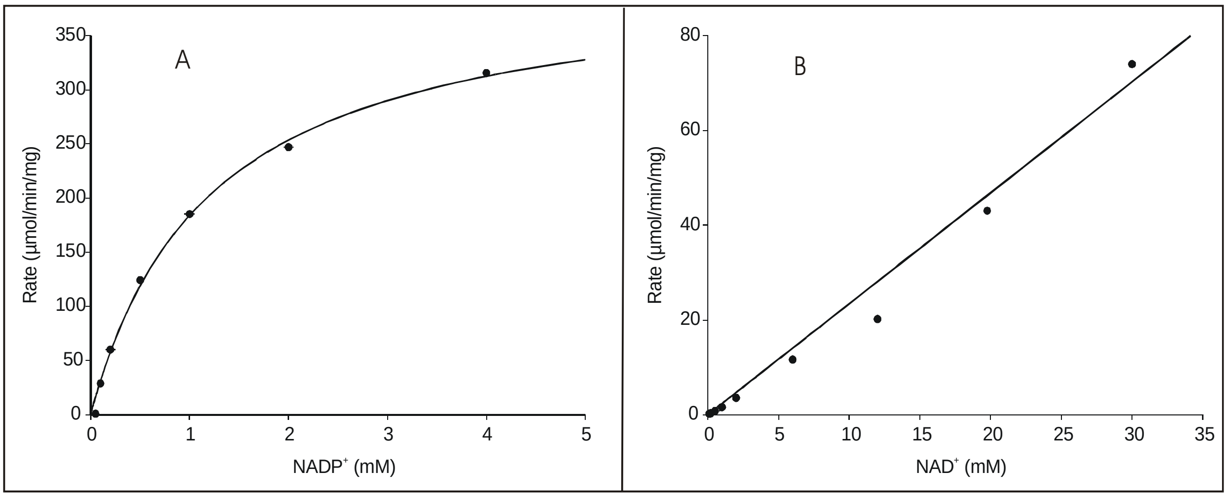


Figure 25. Michaelis–Menten plot of the activity of GdhA with $NADP^+$ (A) or NAD^+ (B) as cosubstrates. The reactions were carried out in the presence of 50 mM D-glucose at 55°C for 10 min in 50 mM phosphate buffer pH 6.5.

3.2.3.3. Temperature, pH optimum and thermoinactivation kinetics

In the standard assay system (50 mM phosphate buffer pH 6.5, 50 mM glucose and 2 mM $NADP^+$, 10 min), the highest rate of glucose oxidation was measured at 55°C. At the optimum growth temperature for *P. torridus* of 60°C, the enzyme displayed 88 % of its maximal activity. The pH optimum of the pure enzyme was determined to be pH 6.5, but at the physiological pH of 4.6 found in the cytoplasm of *Picrophilus* cells it showed merely 10 % of its maximal activity. Additionally, incubation at 60°C (the optimum growth temperature of *P. torridus*) and pH 4.6 in McIlvaine or acetate buffer without supplementation of Zn^{2+} for 1 h led to almost complete loss of enzyme activity. Thermal inactivation kinetic experiments performed at pH 6.5 without the addition of Zn^{2+} to the buffer demonstrated a $t_{1/2}$ of 5 min at 70°C and of 3 h at 65°C (Fig. 26).

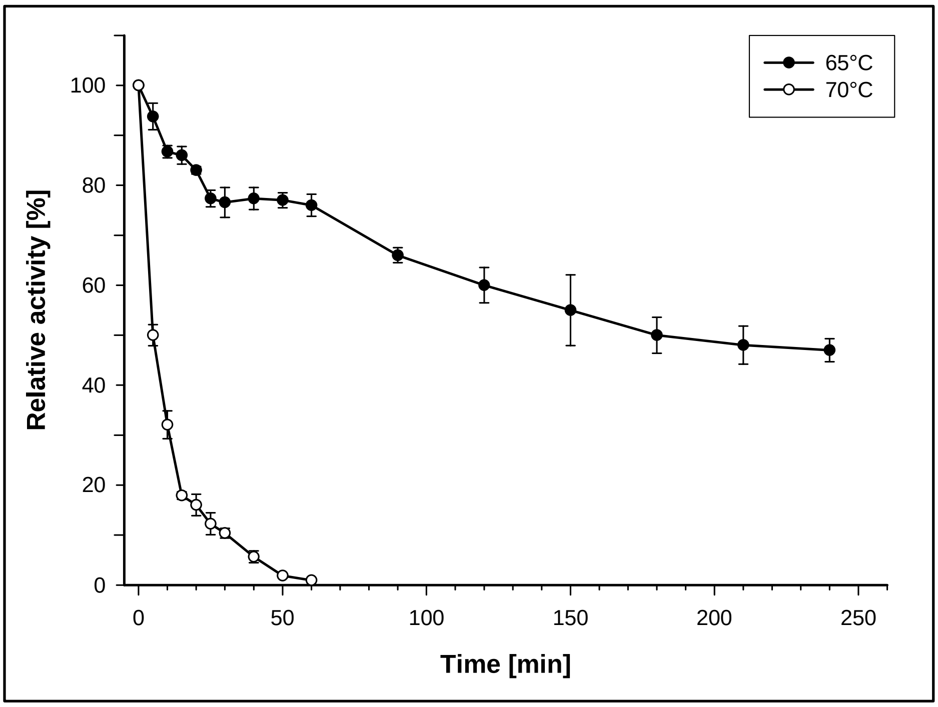


Fig. 26. Thermal inactivation kinetics of the purified recombinant GdhA. The enzyme (0.78 mg/ml) was incubated without the addition of $ZnCl_2$ in 50 mM phosphate buffer pH 6.5 and the residual activity was measured at optimal conditions.

3.2.3.4. Influence of Zn^{2+} on the temperature and pH stability of the enzyme

During the purification of GdhA, a decreasing stability against high temperatures and acidity was observed. Therefore, the influence of factors that could have been lost during the purification process was investigated. Initially, heat treated (2 h, 100°C) *P. torridus* cell extract when added to the incubation buffer was found to recover the stability of GdhA (not shown). Subsequently, this factor was determined to be Zn^{2+}. Addition of $ZnCl_2$ to the assay buffer at up to 5 mM did not show any effect on the activity. Also, no effect was observed with 5 mM NaCl, $MgCl_2$, $MnCl_2$ or $CaCl_2$. EDTA added at up to 10 mM caused no loss of activity. However, the addition of $ZnCl_2$ to the incubation buffers when testing for stability showed a marked effect on the resistance of the enzyme against inactivation at both high temperature and acidity. This effect was the same across the range of $ZnCl_2$ concentrations tested, i.e. from 0.05 to 1 mM. The influence of Zn^{2+} on the pH stability of the enzyme is most evident after incubation (1 h, 55°C) at pH 3.5, where in the presence of the metal ion at 0.1 mM there was 96 % residual activity, opposed to only 5 % in its absence (Fig. 27A). The long–term inactivation kinetics of GdhA at elevated temperatures was also considerably improved by the addition of Zn^{2+}. At 0.1 mM Zn^{2+} incubation at 70°C for 3 h resulted in no appreciable loss of activity.

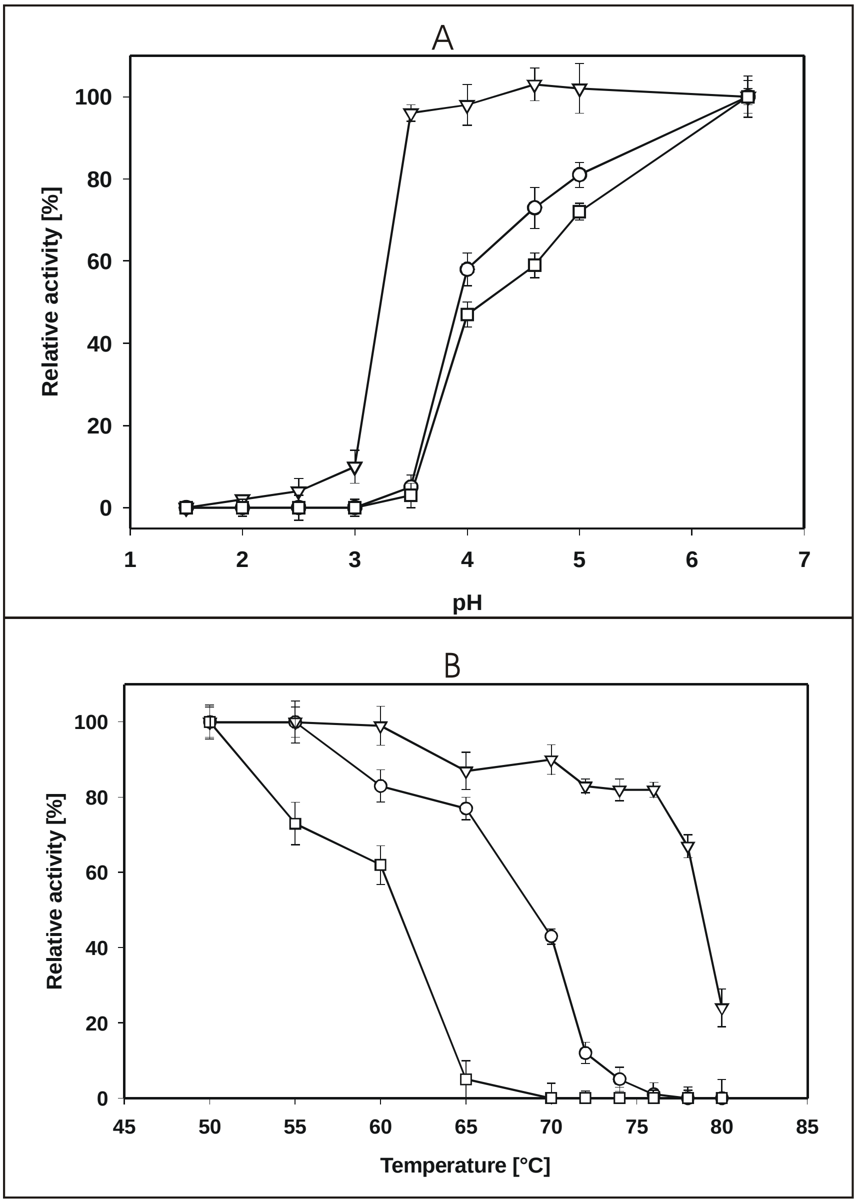


Fig. 27. Temperature and pH stability of GdhA.

A. pH stability
The enzyme (0.12 mg/ml) was incubated for 60 min at the specified acidity at 55°C and the residual activity was measured at standard assay conditions. In accordance with their buffering capacity the following buffers were used at 50 mM concentrations: glycine-HCl, sodium acetate, Tris-HCl, sodium phosphate.

Legend

○ no additive
□ + 10 mM EDTA
▽ + 0.1 mM $ZnCl_2$

B. Temperature stability.
The purified enzyme (0.12 mg/ml) was incubated for 30 min in phosphate buffer pH 6.5 at different temperatures and the remaining activity was measured at optimal conditions. Activity is expressed as [%] of the activity at 50°C.

The specificity for Zn^{2+} in stabilising GdhA was confirmed by incubating the enzyme for 30 min at 75°C in the presence of 1 mM NaCl, $MgCl_2$ or $CaCl_2$, where the remaining activity did not differ from that of the sample incubated in the absence of salts (data not shown). Also, 10 mM EDTA completely abolished the stabilizing effect of Zn^{2+} (Fig. 27).

3.2.3.5. Stability of GdhA in organic solvents

The purified enzyme was considerably stable in the presence of organic solvents: overnight incubation (14 h) at room temperature with 50% v/v of acetone, methanol or ethanol did not result in a detectable loss of activity. In addition, in the presence of 20% ethanol, 30% methanol and 40% acetone in the reaction assay, GdhA still displayed half of its maximal activity (Fig. 28).

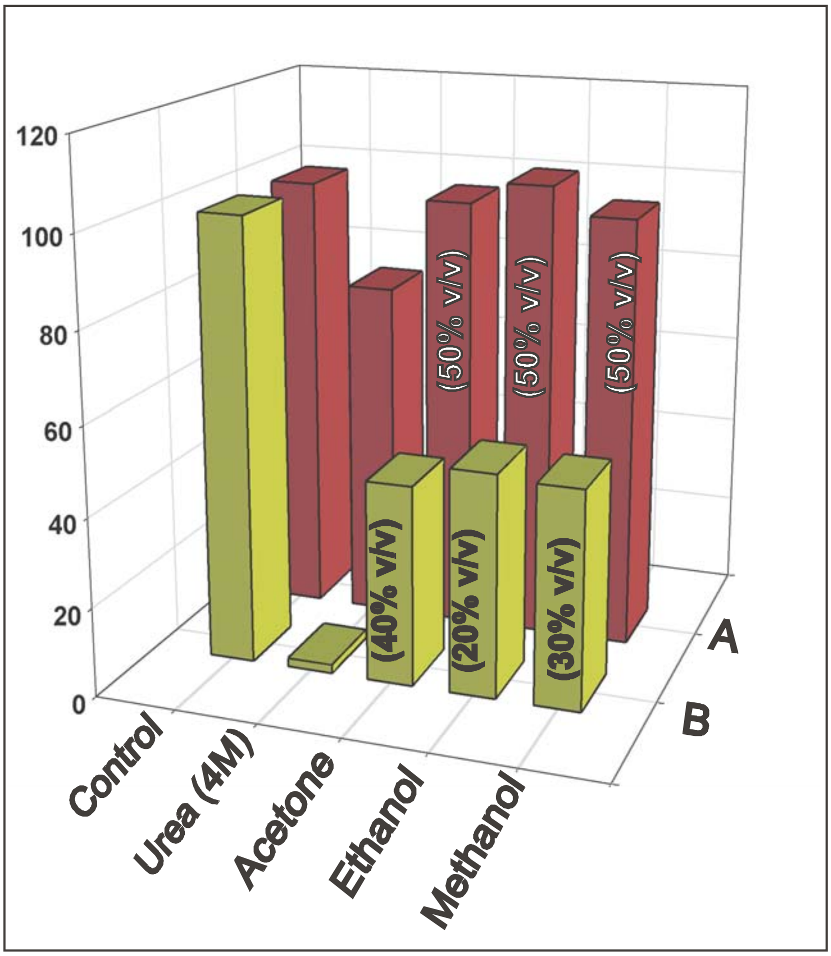


Fig 28. Stability and activity of GdhA in the presence of organic solvents.

A. 0.3 mg/ml purified enzyme was incubated for 14h with the specified compounds at room temperature, and the remaining activity measured in phosphate buffer pH 6.5 and 55°C.

B. The compounds were supplemented in the assay buffer at the specified concentrations and the activity is expressed as % of the control (no additives)

3.2.3.6. Non-enzymatic NADPH degradation

NADPH is rapidly degraded when exposed to high temperature or low pH (Wu *et al.*, 1986). Other reported factors that affect the stability of this compound are the buffer composition, in particular phosphate and acetate, and ionic strength. The chemical hydrolysis of NADPH was followed by measuring the decrease of its absorbance at λ=340 nm (Fig. 29).

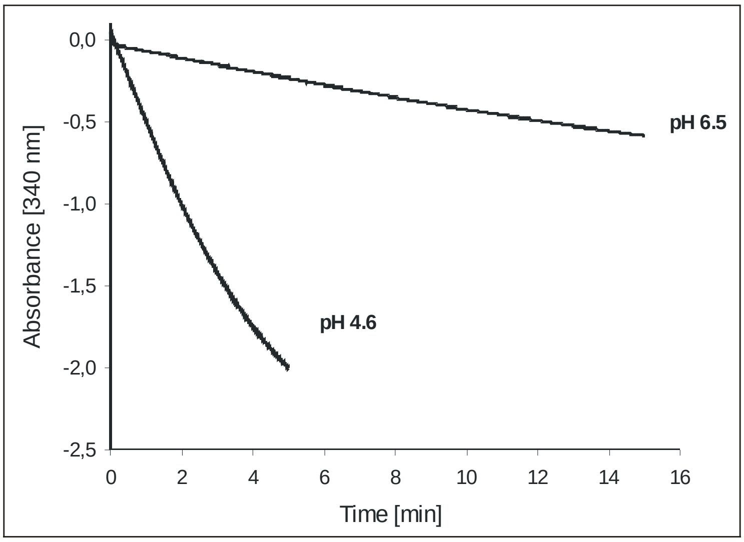


Fig 29. Rate of NADPH degradation at 55°C.

The decrease in absorbance was measured in acetate (pH 4.6) or phosphate (pH 6.5) buffer at 55°C with 2 mM NADPH.

As it is the product whose absorbance is actually measured with the standard assay procedure (see section B.3.6.1), it was important to evaluate the effect of NADPH degradation on the results obtained when it was used. The factor that has the highest impact on the stability of NADPH has been shown to be hydronium ion concentration (Wu *et al.*, 1986) and therefore pH optimum measurements were expected to be biased. When, as an alternative, glucose decrease was used to determine enzyme activity there was a slight shift in the pH at which the GdhA was most active, *i.e.* from 6.5 to 6. Also, the enzyme showed increased relative activity at pH values ranging from 4 to 6 (Fig. 30).

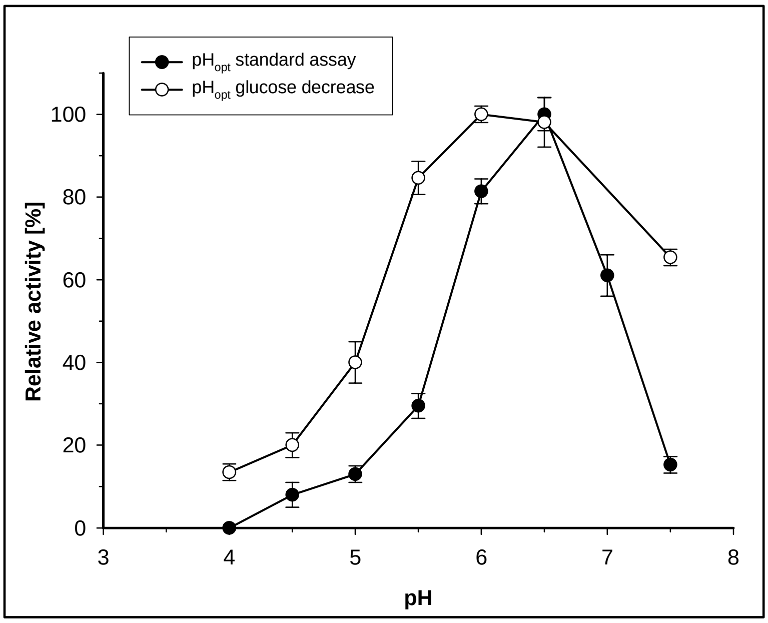


Fig. 30. pH optimum of GdhA. The activity was measured at 55°C by using the standard assay procedure or by determining the rate of decrease of D-glucose (10 min assay).

3.2.3.7. Identification of the native glucose dehydrogenase activity in *P. torridus*

In order to identify the native GdhA in *P. torridus* cells, the pH optimum and temperature optimum for the glucose dehydrogenase activity in crude cellular extracts was determined. Both optima (pH6.5 and 55°C, respectively) were in concert with the optima of the recombinant enzyme. Further evidence in support of the identity of the recombinant enzyme reported here with the enzyme present in *P. torridus* cells is the ratio of enzyme activity with NAD^+ and $NADP^+$ as cosubstrates, which was approximately 1 : 20 in both cases, as well as the ratio of the D-glucose and D-galactose oxidation rates when using $NADP^+$ (Table 15).

Table 15. Comparison of the measured properties of the native *P. torridus* glucose dehydrogenase activity with the recombinantly expressed GdhA.

Parameter	Glucose/galactose dehydrogenase activity in crude *P. torridus* extract	Recombinant GdhA
Temperature optimum	55 °C	55°C
pH optimum	6.5	6.5
$NADP^+$: NAD^+ ratio of glucose oxidation activity	20.1	19.4
D-glucose : D-galactose ratio of dehydrogenase activity	1.43	1.35

Finally, upon native PAGE and subsequent zymogram staining for Gdh activity the recombinant enzyme was indistinguishable from the band of the cell-free *P. torridus* fraction (Fig. 31).

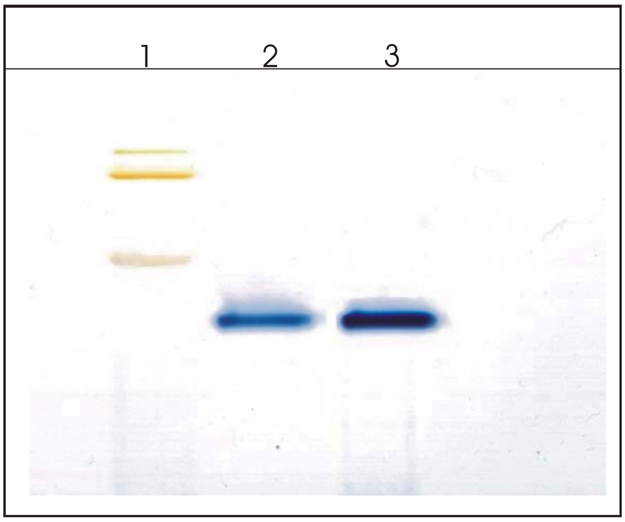


Fig. 31. Native PAGE stained for glucose-dependent $NADP^+$ reduction. The staining buffer contained 1mM $NADP^+$, 50 mM glucose, 1 mM nitro blue tetrazolium, 0.025 mM PMS.

Lane1: MWM containing ferritin (450 kDa), katalase (240 kDa) and cytochrome C (12.5 kDa).

Lane2: Recombinant GdhA (30 mU)

Lane3: *P. torridus* crude extract (30 mU)

D. Discussion

1. The genome sequence of *P. torridus*

1.1. Sequencing and assembly

In the whole-genome shotgun sequencing approach of small genomes a compromise has to be made between the number of the generated shotgun sequences and the time and effort spent in the gap closure phase. From the graphical representation of equation (2) in section C.1.1 it can be seen that there is an "optimal" number of shotgun sequences that would result in a reasonable coverage of a genome of a certain size (Fig. 32).

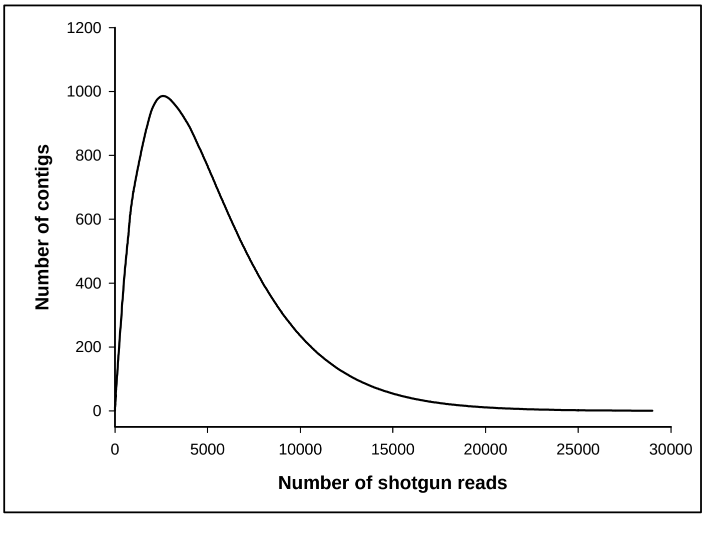


Fig. 32. Graphical representation of the dependency between the number of sequence reads and the number of gaps (resp. contigs) that would be obtained theoretically in a genome shotgun sequencing approach, according to Lander and Waterman (1988). In the calculations, an average sequence length of 600 bp and a genome size of 1.5 Mb were assumed.

Increasing the number of these sequences (respectively the coverage) above a given level would lead to an unacceptably small decrease of the contig number. On the other hand, obtaining a high quality consensus sequence with a coverage below 6 would require a considerable effort for gap closure. In the case of the *P. torridus* genome sequencing project a 8-fold redundancy requiring 20,000 reads was chosen as a strategy that would theoretically result in 6 gaps. There are however several considerations that should be taken into account when interpreting these predictions, i.e., the quality of the shotgun library, the average length and quality of the sequence reads, the presence of repeated regions in the genome etc. Therefore, the 26 gaps obtained in the assembly phase of this project could be considered to be in agreement with the ideal model described above. After the directed sequencing phase, the *P. torridus* genome sequence had a 9.4-fold coverage and an extremely low probability of an error - 1 in 2,000,000. This error rate was achieved by sequencing the regions with the lowest confidence values with custom sequencing primers.

There has been a long lasting discussion on the value of complete genome sequencing of organisms of interest as opposed to the release of draft sequences, in which no gap closing is accomplished (Selkov *et al.*, 2000; Siebers *et al.*, 2004). Some of the advantages of complete genome sequencing are that it allows a comprehensive metabolic pathway reconstruction, comparative genomic studies with closely related organisms and the possibility to detect events in the evolutionary history of the organism like gene losses, duplications or lateral gene transfer. On the contrary, a draft sequence that is composed of a group of contigs whose order and orientation are not known could contain misassemblies and sequencing errors and there might be genome regions not present in the data set. In the *P. torridus* genome project, the aim was to obtain the whole genome sequence of the organism, in which the order and accuracy of every base pair is verified, and thus to take the full advantage of the possibilities discussed above.

1.2. Sequence analysis

With a genome size of 1.54 Mbp, *P. torridus* has the smallest genome of a non-parasitic free-living organism. The genomes of thermophilic methanogenic archaea and hyperthermophilic bacteria, however, are not much larger (about 1.6-1.8 Mb). One of the many effects of high temperature on the organisms living in such environments is an increased error rate in their nucleic acids due to cytosine deamination (Wang *et al.*, 1982). This may have led to a selective pressure in direction of reduced genome size. Although *P. torridus* is a moderate thermophile (T_{opt} 60°C), it must overcome the combination of this temperature and pH values around 0 in the medium and the lowest known intracellular pH (4.6) of all organisms.

Another characteristic feature of the P. torridus genome is the 91.7 % coding sequence. This density is the highest reported for the genomes of thermoacidophilic organisms (89 %, 87 %, 85 %, 85 % for *T. acidophilum, T. volcanii, S. solfataricus* and

S. tokodaii, respectively) and it also seems to be a result of the pressure exerted by the extreme living conditions on the genome organisation.

1.2.1. Amino acid composition and isoelectric point distribution

Due to its unusually low intracellular pH of 4.6, it is expected that the proteins of *P. torridus* could have evolved features, which distinguish them from the proteins of other organisms and which could be detected by analysing their amino acid composition and isoelectric point distribution. However, the amino acids composition analysis of *P. torridus* ORFs and the estimated isoelectric point distribution of the encoded proteins showed no striking differences to the reference organisms (Figures 5 and 6). Only in the case of the halophile *Halobacterium sp*. there was a considerable deviation from the average for both analysed parameters. For the *P. torridus* proteins, only one small deviation was observed, i.e. a slight average increase in their isoleucine content. This is in agreement with the recently published view by Schafer *et al.* (2004) that an increase in hydrophobic residues on the protein surface may be connected with acid stability. Therefore, obtaining 3D structures from *P. torridus* proteins is believed to give insight into the adaptation of proteins to high acidity.

An interesting observation can be made when the isoelectric point distribution of the encoded proteins of *P. torridus* is compared to the one of *Helicobacter pylori*. In the case of the human pathogen which, during its life cycle, has to survive pH values around 2 in the gastric mucous layer, more than 70% of the proteins have an isoelectric point greater than 7.0 compared to 51% for *P. torridus* and 45.3 % for *E. coli*. It has been speculated for *H. pylori* that this distribution reflects adaptation to high acidity (Tomb *et al.*, 1997). This possibly is true for the particular case of *H. pylori* but seems not to be a general feature of acidophiles and may very well be an exception (see Fig. 6).

1.2.2. Phylogenetic analyses

With the genome sequence of *P. torridus*, five complete genomes from thermoacidophilic organisms are available which gives an opportunity to investigate in more detail the evolutionary relationship of organisms within a unique ecological niche. Phylogenetic analysis based on 16S rRNA sequences places *P. torridus* within the order

Thermoplasmales of the euryarchaeal branch of the domain *Archaea* (Fig. 33A). The members of this order are typically aerobic or microaerophilic, heterotrophic organisms which inhabit hot and acid environments. A second group of thermoacidophiles is found in the crenarchaeal branch, the *Sulfolobaceae*. Members of both groups are found to share the same habitats (Johnson *et al.*, 2003) and this is a prerequisite for lateral gene transfer events (Deppenmeier *et al.*, 2002). In the case of *P. torridus* it was observed that:

- Of the 397 ORFs for which no function could be assigned 318 showed similarities to hypothetical ORFs of other organisms. 174 of the latter ORFs have orthologs only in the genomes of other thermoacidophilic organisms, indicating that the thermoacidic environment forms an old and genetically distinct niche. This can be seen also in the observed distribution of the *P. torridus* ORFs when they are presented on the circular chromosome map with indication of their apparent origin – archaeal, bacterial, thermoacidophilic, or unique (Fig. 4).
- Another line of evidence for genetic relatedness within the thermoacidophilic group is the result of a whole-genome comparison for homology on the amino acid sequence level of the complete genomes of three prominent members of this group: *P. torridus*, *T. acidophilum* and the crenarchaeon *S. solfataricus*, (Fig. 33B). When 30% sequence identity was set as a cutoff value for homology, *P. torridus* and *T. acidophilum* showed significant homology in 66 % of all their genes, and these two euryarchaea shared 58 % and 62 % genes respectively with the crenarchaeon *S. solfataricus* but only about 35 % with the phylogenetically more closely related euryarchaeon *P. furiosus*. The assertion that *P. torridus* shares nearly the same number of homologs with *T. acidophilum* and *S. solfataricus* but significantly less homologs with *P. furiosus* remains true even when lowering the threshold for homology from 30 to 25 % identity. Consequently, for these homologous ORFs there is a contradiction between their genealogy and the organismal 16S rRNA tree. This can mean that, at least in this case, ecological closeness overrides phylogenetic relatedness in terms of genome contents.

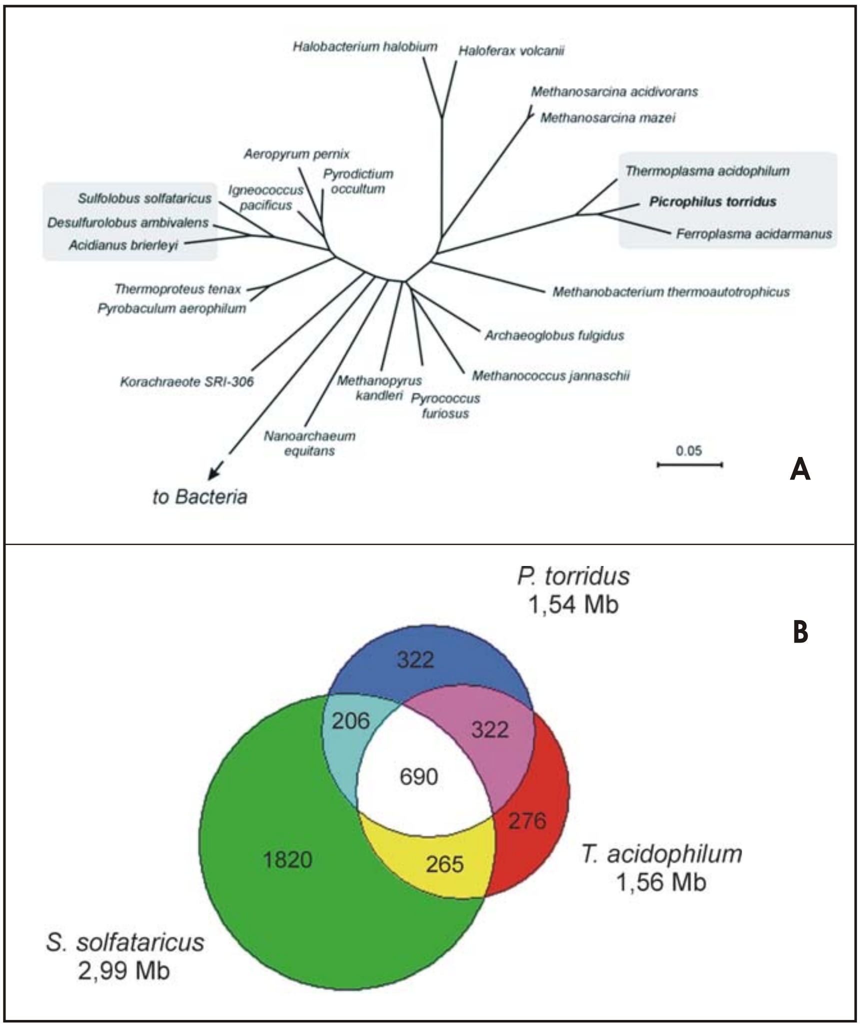


Fig. 33.

A. 16S rRNA phylogenetic tree. Highlighted are the two thermoacidophilic groups of the archaea. Sequences were aligned with the ClustalW algorithm. The tree was built by neighbour joining using the Kimura 2-parameter for distance calculation

B. Number of homologous ORFs in *P. torridus, T. acidophilum* and *S. solfataricus*. The size of the circles is proportional to the genome size.

A hypothesis of a common descent, i.e. the existence of a common ancestor which possessed all homologous ORFs in its genome, fails to explain the above discussed distribution of the homologous ORFs. Alternative explanations can be the following:

1) Hidden paralogy – the appearance of gene duplications in the last common ancestor organism and subsequent loss of one of the paralogs in certain lineages. This is an appealing explanation as it can be expected that in similar habitats the same paralogs would be selected for. In agreement with this view, Kellis and coworkers argued that the yeast *Saccharomyces cerevisiae* arose by an ancient whole-genome duplication followed by massive gene lost and specialisation (Kellis *et al.*, 2004). However, when genes or groups of genes are found to be shared among distantly related taxa, this hypothesis is less convincing.

2) Lateral gene transfer (LGT), i.e., the “movement” of genes or clusters of genes across lineages. This hypothesis has been repeatedly proposed to be valid for a number of

sequenced genomes and is now widely accepted (DeLong, 2000). The possible mechanisms for LGT are transformation, transduction and conjugation. The genomes of organisms that have acquired foreign DNA would be “mosaic” in structure – containing ancestral and foreign regions that could be detected by various traits – base compositions, patterns of codon usage and frequency of di- and trinucleotides for example.

The harsh environment in which *P. torridus* lives presents a possible obstacle for the transfer of DNA between cells. Recently it was shown that DNA exchange can occurr in similar habitats and that this exchange can involve large genome regions (Tyson *et. al.*, 2004). The authors claim to have reconstructed two near-complete genomes (*Leptospirillum* group II and *Ferroplasma* type II) wholly by random shotgun sequencing of a DNA sample derived from a natural acidophilic biofilm. One of the interesting findings obtained by using this approach is the “mosaic” structure of some individual genomes in the *Ferroplasma* type II population. These genomes contained various combinations of genotypes found in the population and undoubtedly show that even in this hostile environment DNA exchange occurs, although among organisms of the same species. Moreover, this genetic exchange was not restricted to small DNA fragments but rather resembled homologous recombination, involving large genome regions. This genome sequencing strategy is highly beneficial in providing valuable information about the gene contents of unculturable microorganisms and accessing the diversity at the genome level within and between natural populations.

1.3. Bioenergetics and central metabolism

In this section, several aspects of the bioenergetics and metabolism of *P. torridus* are discussed in connection with the information obtained from the genome sequence analysis: the transport and utilisation of metabolite substrates, respiration and generation and maintenance of membrane potentials.

1.3.1. The cytoplasmic membrane and the generation of membrane potential

The energy status of the cell is determined by two energy generating processes, energy transduction processes, in which an electrochemical ion gradient is transformed into other forms of energy, and substrate-level phosphorylation processes. For the generation of electrochemical gradients, specific pumps translocate protons or sodium ions across the membrane into the external medium (Speelmans *et al.*, 1993). In the case of proton extrusion, the ensuing electrochemical gradient results in the formation of a proton motive force (PMF). The PMF consists of two components: ΔpH – the concentration gradient of protons across the membrane, and Δψ – the membrane potential which is caused by the transport of the electrical charge of the protons (Mitchell, 1966):

$$PMF = \frac{\Delta\mu_{H^+}}{F} = \Delta\psi - 2.303\frac{RT}{F}\Delta pH \qquad (3)$$

where R is the gas constant, T the absolute temperature (K) and F the Faraday constant. In order to keep the driving force on the protons directed into the cell, the PMF has to be negative. In organisms that live around neutral pH values both components (electrical and concentration) are negative and in this way contributing to the PMF. In the special case of acidophiles, a large ΔpH exists, in the case of *P. torridus* 4-5 pH units, which could eventually lead to H^+ influx and acidification of the cytosol. This can be overcome by several possible strategies: Reversing the Δψ to positive inside, decreasing the permeability of the membrane for protons, and decreasing the internal pH would all help to neutralise the high ΔpH. Indeed, in most acidophiles as well as in *Picrophilus oshimae*, a member of the *Picrophilaceae* family, the Δψ is positive inside, extremely low permeability of the cytoplasmic membrane has been reported, and an internal pH of 4.6 has been measured when the pH of the medium varied from 0.7 to 4 (van de Vossenberg *et al.*, 1998). This is lower than the values reported for other thermoacidophilic organisms, which maintain a pH around neutral (She *et al.*, 2001).

1.3.1.1. Cell wall and cytoplasm membrane

One of the major keys to the adaptation of *P. torridus* to the acidic environment is the nature of its cell wall and membrane. The membranes of *Picrophilus sp.* mainly consist of polar ether lipids like caldarchaeol or modified derivates thereof and are highly impermeable to protons (van de Vossenberg *et al.*, 1998). The cell wall is formed by an S-layer with tetragonal symmetry and center-to-center distance of about 20 nm, attached with thin pillars to the membrane. On its outer surface, brush-like structures have been observed, possibly made of long polysaccharide chains (Schleper *et al.*, 1995). Several ORFs with a probable role in diether and tetraether lipid biosynthesis and a putative S-layer protein gene were detected. However, it was not possible to deduce reasons for the membrane or cell wall acid resistance by the genome sequence alone. More information, e.g. about the physico-chemical properties and structural features of the S-layer proteins are needed to gain a better understanding of the acid resistance of the *P. torridus* cell surface.

1.3.1.2. Respiration

Obligate aerobes are not commonly found among Archaea (Schäfer *et al.*, 1999). In contrast to its close relative, the microaerophilic *T. acidophilum, P. torridus* is an obligate aerobe and uses a more complex electron transport chain in order to maintain the membrane potential. The genome data shows that all type I NADH oxidoreductase-homologous genes of *Paracoccus denitrificans* are found in *P. torridus* except the ones coding for the input module, *nuoEFG* (Fig. 9). The same is true for the complex I of *T. acidophilum* and *F. acidarmanus* but not for members of the *Sulfolobales* which lack some of the integral membrane and electron transfer subunits (Ruepp *et al.*, 2000, She *et al.*, 2001). Therefore, it can be assumed that complex I in thermoacidophiles of the euryarchaeota is able to transfer protons over the cytoplasmic membrane in contrast to complex I of the thermoacidophiles of the crenarchaeota. It is interesting to note that in most archaea no genes for homologs of the subunits of the electron input module which oxidises NADH and subsequently channels the electrons to the membrane-associated quinone reductase module can be found. It is still unknown how electrons are fed into the transport chain in organisms without the NADH input module and whether NAD^+ or ferredoxin is the electron mediator between metabolism and the electron transport chain (Schäfer *et al.*, 1999; Deppenmeier *et al.*, 2002).

Succinate deydrogenases (SDHs) have been found in the three domains of life. In the archaeal domain, two different SDHs are described, classical and non-classical ones (Schäfer *et al.*, 1999). The latter possesses unusual membrane-associated subunits (sdhC/D) with similarities to membrane-bound subunits of the heterodisulfide reductase of methanogenic archaea and are usually found only in crenarchaeal thermoacidophiles but also in the genomes of *P. torridus* and *F. acidarmanus*. It can be speculated that *P.*

torridus has acquired the complete *sdh* operon from a crenarchaeon since in *T. acidophilum* and all other non-thermoacidophilic archaeal genomes only classical SDHs are found.

Since several components of the respiratory chain of *P. torridus* are by far most similar to genes from organisms of the distant crenarchaea or bacteria but are not found in members of the *Thermoplasmaceae*, it seems likely that these were obtained relatively late in evolution by horizontal gene transfer.

1.3.2. Transporters

Besides a K^+-channel, *P. torridus* seems to possess a K^+-transporting ATPase. It has been proposed for the genome of *S. solfataricus* (She *et al.*, 2001) that its functional role is most probably to invert the $\Delta\psi$ to positive inside by potassium uptake in order to cope with the high ΔpH (see section D.1.3.1). In the case of *P. torridus* this role is even more probable, due to the extremely high ΔpH that has to be decreased.

Interestingly, the overall ratio of genes involved in secondary transport to primary transport is 5.6:1 which is unusually high compared to other microorganisms such as *S. solfataricus* (2.7:1), *E. coli* (2.6:1), *Pyrococcus horikoshii* (1.5:1) or *Thermotoga maritima* (0.5:1). As no candidate genes for secondary transporters which use Na^+ to drive the transport could be found, it can be concluded that *P. torridus* relies mainly on the high proton motive force to drive its metabolite transport. The high number of ABC-transporter genes for peptide and sugar uptake on the other hand indicates the importance of such compounds as nutrient sources, and points to the need of high-affinity transporter systems for the efficient uptake of these substrates.

Importantly, the exceptionally high ratio of secondary to primary solute transport systems found in *Picrophilus* indicates that the predominant use of proton-driven secondary transport represents a highly relevant strategy for the adaptation of this organism to its extremely acidic environment. In contrast, it is known that in most hyperthermophilic bacteria and archaea primary uptake systems are preferred (Albers *et al.*, 2001). This strategy of a high secondary-to-primary transporter ratio for acidophilic adaptation can also be observed in the genomes of other thermoacidophilic euryarchaea like *T. acidophilum* and *F. acidarmanus* but not in thermoacidophilic organisms of the crenarchaeota and thus seems to be a trait only common to the former branch.

1.3.3. Carbohydrate metabolism

P. torridus and *P. oshimae* produce extracellular thermo- and acid stable glucoamylases (Serour *et al.*, 2003), meaning that in the case of α-glucan substrates extracellular enzymatic breakdown can occur. In addition, it seems clear that sugar polymers outside the cells of *P. torridus* could be degraded also, at least partially, by the acidic and hot environment. Inside the cell, oligomeric sugars can be hydrolysed by two predicted intracellular glucoamylases, an α-amylase and an α-glucosidase. The gene coding for one of the glucoamylases has recently been cloned in *E. coli* and the function of the encoded enzyme confirmed (B. Schepers, personal communication). Also, as a part of the current work, the α-glucosidase gene (*mal*P, ORF 985) was expressed in *E. coli* and the activity of the encoded product was shown to be consistent with the annotation (section C.3.1.3).

Despite the fact that the presence of glycogen has been shown in several archaeal species, little is known about its metabolism in this group (König *et al.*, 1982). From genome data alone it is possible only to speculate about the probable role of the ORFs thought to be involved in glycogen turnover. In yeast and mammals the hydrolytic degradation of glycogen is accomplished by a debranching system composed of a polypeptide chain that contains two activities: 6-α-glucosidase and 4-α-glucanotransferase (Bates *et al.*, 1975). The 4-α-glucanotransferase activity acts on the glycogen phosphorylase limit dextrin chains to expose the single glucose residues, which the 6-α-glucosidase activity can then hydrolyse. In *P. torridus*, the intracellular α-amylase (ORF 596) was found in a potential operon of functionally related genes, which is conserved in the *Thermoplasmales*. Unexpectedly, in *P. torridus* the 6-phosphofructokinase gene was located upstream of this conserved region. The cluster consists of ORFs for glycogen debranching enzyme (GDE), a glycosyl transferase (GT) and the α-amylase (AmyA). The overlap between the *P. torridus* ORFs for GDE and GT was found to be 2bp and between those for GT and AmyA – 16 bp. It is interesting to point out that the deduced GDE amino acid sequence showed significant sequence similarity to the N-terminal part of the yeast glycogen debranching enzyme, which is where the 6-α-glucosidase activity is located. The conserved probable operon in several members of the *Thermoplasmales* and the supposed mode of action of the encoded enzymes on glycogen is shown on Figure 34.

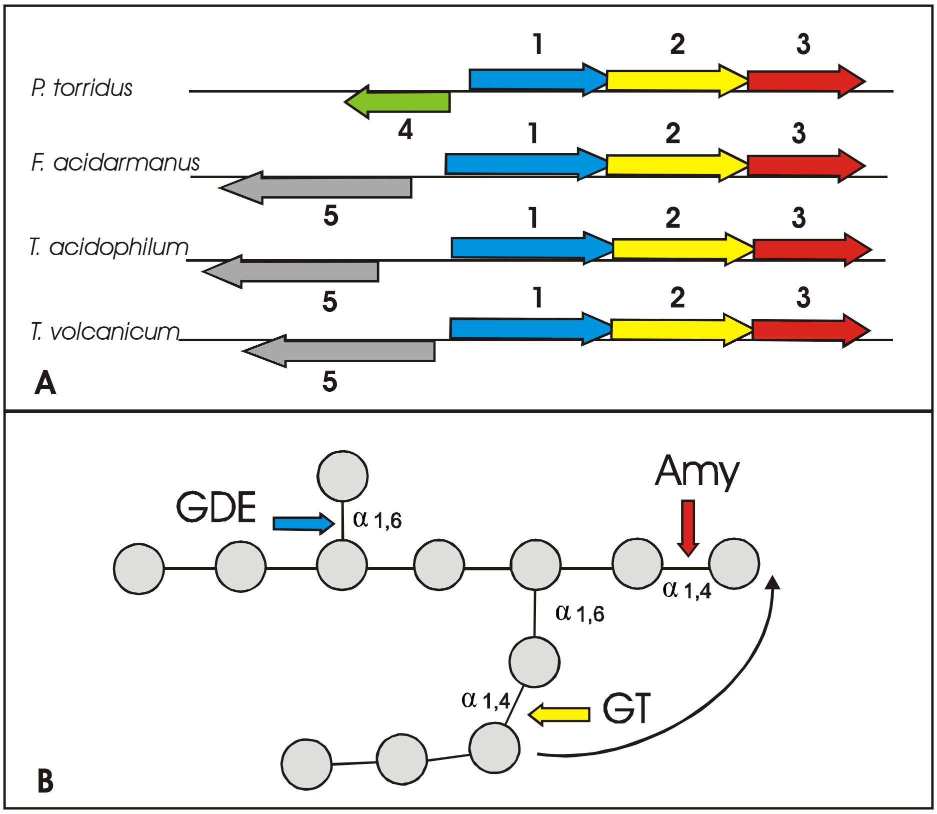


Fig. 34. A. Organisation of the genes functionally related to glycogen metabolism in *P. torridus, F. acidarmanus, T. acidophilum* and *T. volcanicum*. The numbers correspond to ORFs for: 1. glycogen debranching enzyme (GDE), 2. glycosil transferase (GT), 3. α-amylase (AmyA), 4. 6-phosphofructokinase and 5. glucoamylase.

B. Mode of glycogen degradation by the enzymes encoded in the cluster. The circles represent glucose residues

As has been described in the Results (section C.1.3.8), *P. torridus* most likely catabolises glucose via a non-phosphorylated variant of the Entner-Doudoroff (ED) pathway. The identification of the gluconate dehydratase gene, which previously had not been identified in other archaeal genomes was based on strong similarities to galactonate dehydratase genes of enterobacteria. In support, orthologs of this ORF are clustered with the KDG aldolase gene in *S. solfataricus* and with the glucose dehydrogenase gene in *T. acidophilum*.

Concerning the EMP pathway, it was assumed that it is functional despite the lack of a candidate gene for a fructose-1,6-bisphosphate aldolase. It has been argued previously that alternative enzymes are utilised to obtain the missing activity in *T. acidophilum* (Ruepp *et al*., 2000). Interestingly, *P. torridus* possesses, in contrast to *T. acidophilum* and *S. solfataricus*, a phosphofructokinase gene which would be unnecessary unless its reaction product is further cleaved in an aldolase reaction or vice versa in the gluconeogenic orientation of the pathway. Recently, Lamble and coworkers

(Lamble *et al.*, 2003) revealed the promiscuity of the KDG-aldolase of *S. solfataricus* which in its reverse reaction synthesises 2-keto-3-deoxygalactonate besides 2-keto-3-deoxygluconate from pyruvate and glyceraldehyde. Since also the glucose dehydrogenase of this organism does not differentiate between glucose and galactose, it was assumed that this is true for the complete pathway which in consequence would argue against the use of this pathway (ED) for gluconeogenesis. The biochemical data obtained for the *P. torridus* glucose dehydrogenase, i.e. its high enzyme activity with galactose as substrate and its high affinity for this sugar (approximately double compared to glucose), suggest that this “promiscuity” is probably present also in *P. torridus*. It is therefore proposed by us that a non-classical fructose-1,6-bisphosphate aldolase may be present in *P. torridus* and that the EMP pathway is used, at least, for gluconeogenesis. The possibility of utilising the EMP for gluconeogenesis is further emphasized by the presence of a putative fructose-1,6-bisphosphatase gene in *P. torridus*. This hypothesis is further discussed in the context of the biochemical characteristics of the recombinant *P. torridus* glucose dehydrogenase. (section D.4.1).

It was shown in the Results that genes coding for all enzyme components of complete TCA and 2-methylcitrate cycles could be detected in the *P. torridus* genome (section C.1.3.9). Interestingly, it was reported previously that the addition of propionate, lactate, acetate or formate to *P. oshimae* cells inhibited respiration (van de Vossenberg *et al.*, 1998). Since also a lactate-2-monooxygenase was found which converts lactate to pyruvate, two acetyl-CoA synthetases, parts of a formate hydrogen lyase operon and a formyl-tetrahydrofolate synthetase, it is possible that the tested compounds are not metabolised in substantial amounts but the enzymes and pathways serve mainly as a means of protection against uncoupling of the respiratory chain by organic acids. This uncoupling is probable to occur because at the pH at which *P. torridus* grows (0.6-1.0) these weak organic acids are protonated and therefore able to enter the cell, where by deprotonation at the higher pH (4.6) they would “work against” the respiration.

It must also be noted that the formyl-tetrahydrofolate synthase ORF shows high similarity to bacterial genes and a lactate monooxygenase homolog has so far not been detected in any other archaeal genome. It can therefore be concluded that, for at least some of the organic acid metabolic pathways, their presence in the *P. torridus* genome is most probably due to horizontal gene transfer.

1.3.4. Overview

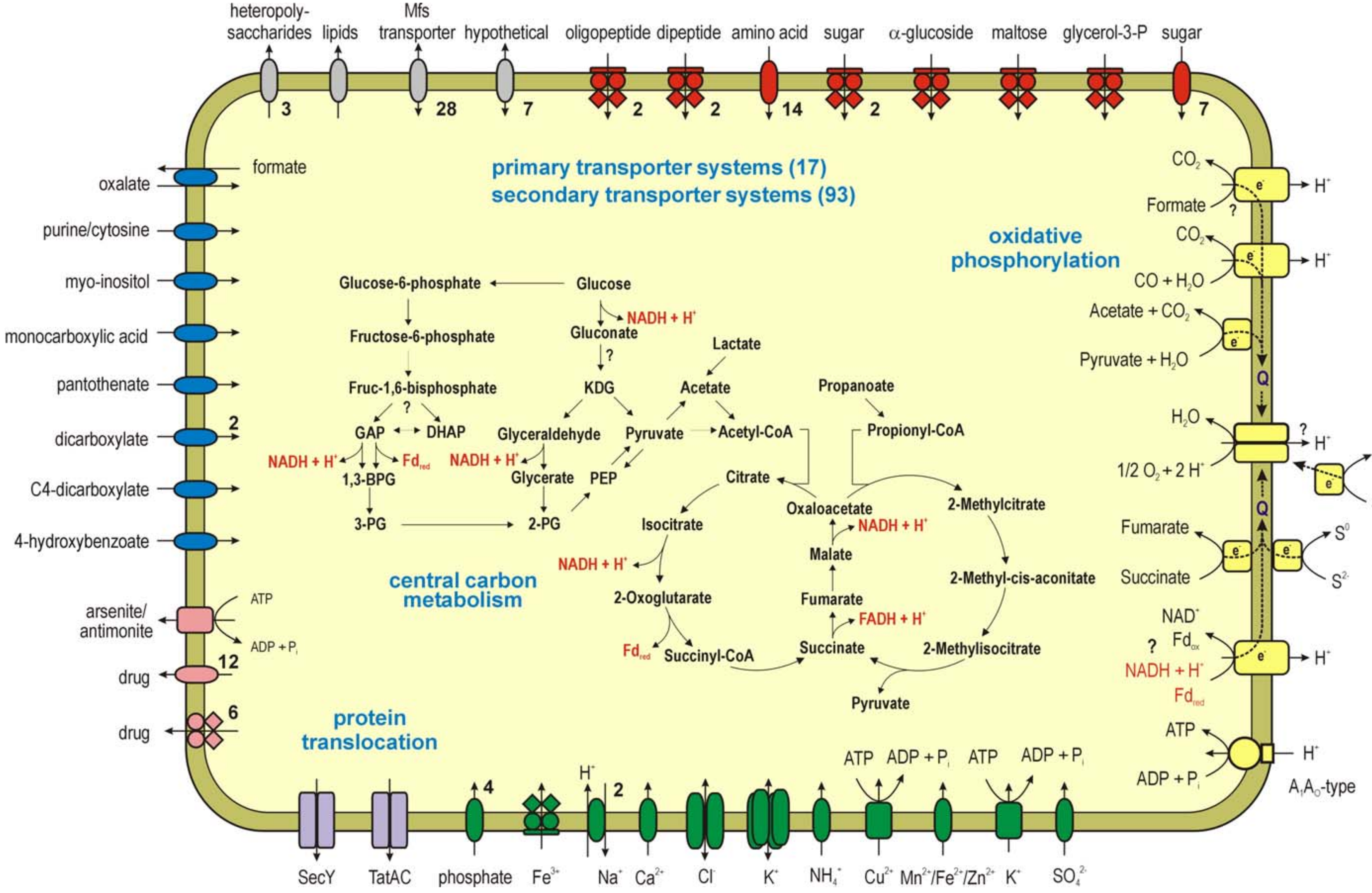
It is important to note that many genes which are connected with the abilities of *P. torridus* to survive in its extremely acidic environment have been obtained by lateral gene transfer. This includes some of the organic acid degradation pathways, the main components of the electron transport chain and mechanisms to deal with oxygen stress.

Below is an overview of the transport, central metabolism and energy production in *P. torridus* as deduced from the annotation of the ORFs derived from the genome sequence.

Fig. 35. Overview of the transport, central metabolism and energy production in *P. torridus*.

Sugar, peptide and amino acid uptake systems are shown in red, drug exporters in pink, trace elements transport systems in green, other and hypothetical transporters in grey. Bold numbers mark the number of each transporter. Protein translocation systems are shown in violet and the components of the respiratory chain in yellow. A total of 93 secondary and 17 primary transporters were found in the genome sequence resulting in an unusual ratio of 5,6:1. So far, no aldolase gene is found. Enzyme activity essays indicate a functional non-physphorylated Entner Doudoroff pathway for glycolysis. Pathways for the respiration of the organic acids acetate, lactate and propanoate were identified. $NADH_2$ and reduced ferredoxin is produced in the *P. torridus* central carbon metabolism but the final reducing compound of the NADH-oxidoreductase is still unknown as no electron-input module for it was detected.

.


heteropoly-saccharides
lipids
Mfs transporter
hypothetical
oligopeptide
dipeptide
amino acid
sugar
α-glucoside
maltose
glycerol-3-P
sugar
3
28
7
2
2
14
2
7
primary transporter systems (17)
secondary transporter systems (93)
formate
oxalate
purine/cytosine
myo-inositol
monocarboxylic acid
pantothenate
dicarboxylate
2
C4-dicarboxylate
4-hydroxybenzoate
arsenite/antimonite
ATP
ADP + P$_i$
12
drug
6
drug
central carbon metabolism
protein translocation
oxidative phosphorylation
Glucose-6-phosphate
Glucose
NADH + H$^+$
Gluconate
?
Fructose-6-phosphate
KDG
Lactate
Acetate
Fruc-1,6-bisphosphate
?
GAP
DHAP
Glyceraldehyde
Pyruvate
Acetyl-CoA
NADH + H$^+$
Fd$_{red}$
NADH + H$^+$
1,3-BPG
Glycerate
PEP
3-PG
2-PG
Propanoate
Propionyl-CoA
Citrate
Oxaloacetate
2-Methylcitrate
Isocitrate
NADH + H$^+$
Malate
NADH + H$^+$
2-Methyl-cis-aconitate
2-Oxoglutarate
Fumarate
FADH + H$^+$
Fd$_{red}$
Succinyl-CoA
Succinate
2-Methylisocitrate
Pyruvate
CO$_2$
Formate
?
e$^-$
H$^+$
CO$_2$
CO + H$_2$O
e$^-$
H$^+$
Acetate + CO$_2$
Pyruvate + H$_2$O
e$^-$
Q
H$_2$O
1/2 O$_2$ + 2 H$^+$
?
H$^+$
S^0
S^{2-}
e$^-$
Q
Fumarate
Succinate
e$^-$
e$^-$
S^0
S^{2-}
NAD$^+$
Fd$_{ox}$
?
NADH + H$^+$
Fd$_{red}$
e$^-$
H$^+$
ATP
ADP + P$_i$
H$^+$
A$_1$A$_0$-type
SecY
TatAC
4
phosphate
Fe^{3+}
H$^+$
2
Na$^+$
Ca^{2+}
Cl$^-$
K$^+$
NH$_4^+$
ATP
ADP + P$_i$
Cu^{2+}
Mn^{2+}/Fe^{2+}/Zn^{2+}
ATP
ADP + P$_i$
K$^+$
SO$_4^{2-}$

2. Heterologous expression of *P. torridus* genes

P. torridus lives in an extremely hostile environment and is able to grow at the lowest pH values known for all organisms. It can be expected that these conditions have led to the selection of distinctive changes in the physiology and metabolism of the organism as well as in the structure and function of its biomolecules. Therefore, one of the aims of the current work was to obtain certain enzymes of *P. torridus* by heterologous expression in order to study their structural and functional characteristics with the ultimate goal to learn how this organism copes with its hot and acid environment.

2.1. Codon usage

Initial cloning and expression experiments with *P. torridus* genes in *E. coli* using the T7-based expression vector pET24c in a BL21 genetic background showed no detectable level of expression (not shown). The reason for such a result most often lies in major discrepancies between the codon usage in the expression host and the introduced protein - it is well known that the codon usage of *E. coli* is highly biased. In particular, arginine AGA and AGG codons are extremely rare, which inevitably affects the heterologous expression of archaeal proteins, where these are the major codons for arginine (Appendix B). And indeed, supplying minor arginine tRNAs in the expression host dramatically improved the expression levels of all the recombinant proteins tested (section C.2.1).

2.2. Inclusion body formation

When overexpressing recombinant proteins, inclusion bodies can be observed in different host systems, for example, prokaryotes, yeast or higher eukaryotes. It is interesting to note that even endogenous proteins, when overexpressed, can result in inclusion body formation (Gribskov *et al.*, 1983). This is an indication that high expression rates are the primary reason for the aggregation of proteins, regardless of the expression system or the host used. On the other hand, proteins with a high hydrophobic content are more predisposed to aggregation due to increased intermolecular interactions of regions of the folding polypeptide chain (Hartl *et al.*, 2002).

The majority of the *P. torridus* proteins whose expression was attempted in *E. coli* with T7 promoter based vectors were accumulated as inclusion bodies (Table 10). Two general approaches were applied in attempting to overcome this problem – (i) purification of the obtained inclusion bodies and subsequent refolding or (ii) escaping their formation. The second approach included the use of different growth conditions in order to decrease the rate of protein synthesis, employment of different expression system (weak promoter) or the construction of fusion proteins with a highly hydrophilic and soluble partner protein. Additionally, coexpression of one or more of the *E. coli* heat shock proteins – GroEL, GroES, DnaK,DnaJ and GrpE was tested. These proteins, although constitutively expressed in *E. coli*, are synthesised at increased levels under stress conditions and are considered responsible for the proper folding of nascent polypeptide chains (Gething *et al.*, 1992).

As an alternative to *E. coli* as the host for expression experiments, expression in *S. solfataricus* using the viral vector SSV1 and in the yeast *S. cerevisiae* for one *P. torridus* protein was tested (sections C.2.4 and C.2.5).

The outcome of using all of the refolding techniques described in chapter C.2.1. was a soluble but inactive enzyme. These soluble forms of the proteins were unstable and easily denatured at moderate temperatures (data not shown), indicating non-native protein structures. It has to be noted, however, that this approach is highly empirical and laborious. Therefore, it is probable that untested conditions exist which would lead to obtaining a particular *P. torridus* enzyme in its functional, native state after refolding. The use of the commercially available kits also led to inactive proteins, at least for the two enzymes used for screening – the two ORFs annotated as β-galactosidase (ORFs 810 and 615).

When the ORF 810 coding sequence was fused in-frame with the *E. coli* genes *nus* or *mal* the resulting polypeptides were expressed in a soluble form (section C.2.3). The same result, i.e. soluble protein, was obtained when the gluconate dehydratase ORF 1383 was expressed as a fusion with the *E. coli mal* gene (Fig. 15). It is obvious that the presence of the N- terminal fusion peptide sequences are responsible for the different processing of these proteins in *E. coli*, since the same ORF expressed without a tag resulted exclusively in inclusion body formation. This effect was dependent neither on the promoter used nor on the genetic background of the host. This can be deduced from the facts that the Nus Tag system is actually based on a modified pET vector and that the same *E. coli* Rosetta strain was used for the expression from both p24-810 and pN-810. On the other hand, when the two fusion constructs of ORF 810 are compared, a different level of solubility of the encoded polypeptides could be observed. The use of the P_{TAC} promoter in the expression vector pM-810 (maltose-binding protein fusion) resulted in a greater fraction of soluble protein compared to the fraction obtained by the T7-based pN-810 (NusA fusion, Figures 16 and 17). This difference can be attributed either to the promoter or to the fusion partner. As NusA has been reported to be one of the most soluble proteins in *E. coli* (Harrison *et al.*, 2000) it is more probable that the use of the weaker P_{TAC} promoter was of greater impact on soluble fusion protein production than the fusion partner. The fusion proteins, however, were also inactive, both before and after the cleavage of the fusion partner. Similar to the results from the refolding experiments, soluble forms were produced which were unstable and easily precipitated after heat treatment at 60°C (Fig. 14B).

The coexpression of the *E. coli* chaperones encoded by the vector pG-KJE8 (Nishihara *et al.*, 1998) also did not result in an active form of the tested protein β-galactosidase encoded by ORF 810 (data not shown).

Because β-galactosidase activity was repeatedly detected in *P. torridus* cell extracts (not shown) it is interesting to compare the amenability to overexpression between the *P. torridus* ORF 810 product and its ortholog in *S. solfataricus* (LacS). The two polypeptides share 52% amino acid sequence similarity and produce reciprocal best BLAST hits. However, LacS from *S. solfataricus* has been expressed in an active form in *E. coli* 5α, using a pUC19-based vector (Haseltine *et al.*, 1999). Also, when using the SSV1 based expression system, the control vector (containing *lacS*) did complement the Lac^- phenotype of *S. solfataricus* PH1-16, while the vector containing the *P. torridus 810* gene did not, as inferred by measuring the β-galactosidase activity in the primary transformation mixture (data not shown). The inability of the ORF 810 construct to complement the Lac^- phenotype in *S. solfataricus* is most probably due to lack of expression. Of course the possibility of a wrong annotation of ORF 810 also exists. The other candidate gene that could account for the measured native β-galactosidase activity is ORF 615 whose translation product shares 22 % amino acid sequence similarity with both LacS and the ORF 810-encoded polypeptide. The results from the expression experiments with this ORF, however, were similar to the ones obtained with ORF 810.

Despite the overall amino acid sequence similarity between *S. solfataricus* LacS and *P. torridus* ORF 810 polypeptide, an important difference can be observed when the hydrophobicity plots of the two encoded proteins are compared. As it can be seen from Figure 36, they have a similar distribution pattern along the polypeptide chain, except for a highly hydrophobic N-terminal region in the *P. torridus* protein.

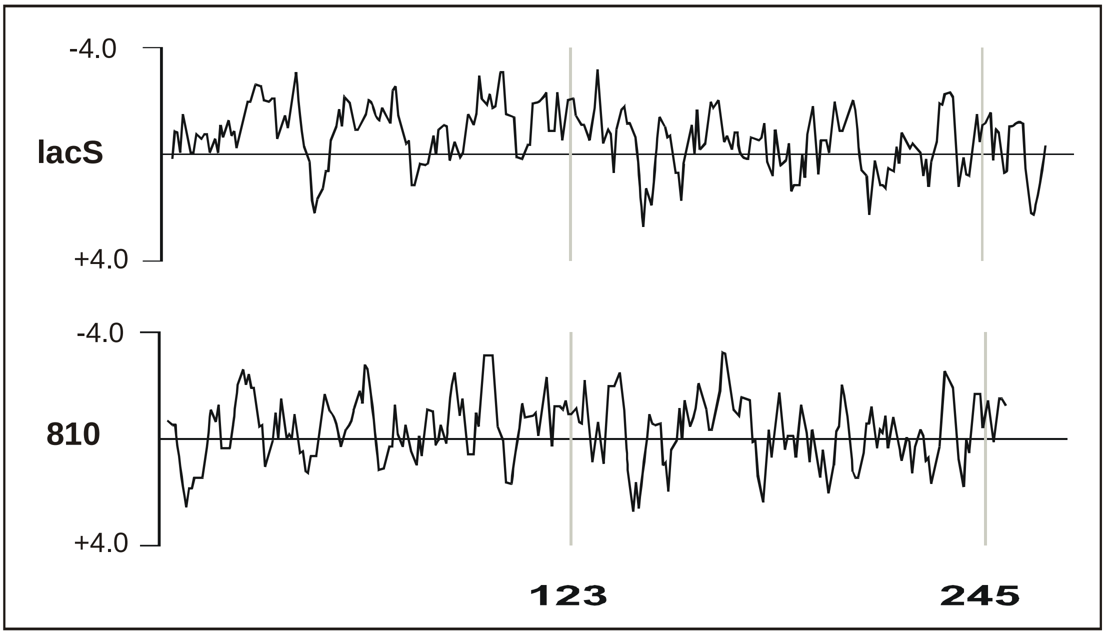


Fig. 36. Hydrophobicity analysis of the N-terminal 250 amino acids of *S. solfataricus* LacS and the predicted *P. torridus* ORF 810 β-galactosidase. The abscissa represents Kyte-Doolottle values, hydrophilic residues have a negative score. Each value is the average of the values of 5 adjacent residues and is plotted at the middle residue.

This hydrophobic N-terminus may be the decisive factor that contributes to the different processing of the proteins compared above both in *E. coli* and in *S. solfataricus*. This is supported also by the successful expression of soluble N-terminal fusion constructs of ORF 810 polypeptide.

Having in mind the low intracellular pH of *P. torridus*, one reason for the failure to obtain properly folded polypeptides in these experiments could be the neutral cytoplasmic pH of the different expression hosts. Further possible reasons include the requirement of specific chaperones (at least when *E. coli* was used), a high hydrophobic content of the tested proteins, or a combination of the above.

2.3. Overexpression of *P. torridus gdhA* and *malP*

A successful outcome of the use of a weaker promoter than T7 was observed in the case of the *P. torridus* glucose dehydrogenase (GdhA, ORF 421, see section C.2.2). When a T7 promoter-based expression vector was used (p24-GDH), a large proportion of the *P. torridus* protein was found as inclusion bodies. Placing the *gdhA* gene under the control of the *araB* promoter in pBAD-GDH allowed optimisation of the expression in *E. coli*, and a substantial amount of active glucose dehydrogenase was obtained. The abundance of rare codons in the GdhA ORF, for example the arginine codons AGG and AGA together accounted for 4.4 % of the total codons, imposed the use of *E. coli* Rosetta as an expression host. Despite the fact that this strain lacks the mutations preventing arabinose catabolism (*araBACD*$^+$) and in this way maintaining constant amounts of inducer in the host is not possible, it is capable of transporting arabinose inside the cell (*araEFGH*$^+$). Therefore, high concentrations of inducer were needed, i.e. 0.2 %, in order to obtain sufficient level of expression. Attempts to transform the plasmid coding for the rare tRNAs (pRARE, Novagen) into *E. coli* Top10 (*araBACD*$^-$) in order to more precisely regulate the transcription from the *ara* promoter were unsuccessful due to instability of this plasmid in *E. coli* Top10 (data not shown).

Expression of the functional active α-glucosidase enzyme (MalP) from *P. torridus* was achieved when the *malP* gene (ORF 985) was cloned in the pET24d vector and *E. coli* Rosetta was used as an expression host (see section C.3.1.2). In this case, active protein was obtained despite the use of the strong T7 promoter and inclusion bodies formation was not observed. However, as indicated by SDS PAGE analysis and by the results obtained from the purification, it is obvious that MalP is produced in this system at an atypically low level. The purification factor of 492 at approximately 37 % yield emphasises the low rate with which the enzyme had been synthesised (Table 12). One of the possible reasons that could explain this observation is the presence of an alternative start codon which was identified 708 bp downstream of the original translation start and which had a consensus *E. coli* ribosomal biding site (...AAGGAG...) at –12 bp. It can be concluded that further optimisation of the expression of *malP* is needed.

In conclusion, the results from the expression experiments indicate that the rate of protein synthesis had a major effect on the state of the proteins that were obtained. This observation is confirmed by the expression of both GdhA and MalP as native proteins in *E. coli* at conditions leading to decreased rates of protein production. On the other hand, this can not be considered as the only factor that contributes to the production of soluble heterologously expressed proteins. Other factors with significant influence were the presence of a N-terminal highly soluble protein fusion partner, and the hydrophobic content of the polypeptide.

3. Properties of GdhA and MalP

3.1. *P. torridus* α-glucosidase (MalP)

Alpha-glucosidases (EC 3.2.1.20) hydrolyse terminal α-D-glucose residues with the release of α-D-glucose, usually from the non-reducing ends of oligosaccharides. The α-glucosidase enzyme from *P. torridus* (ORF 985, designated as MalP) was assigned as a member of the glycoside hydrolase family 31 (CAZy database; Henrissat, 1991). Proteins with significant amino acid sequence similarity to MalP could be found predominantly in the crenarchaeal branch of the *Archaea* and in the bacterial and eukaryal domains. No organisms closely related to *P. torridus* seem to contain a similar enzyme. Interestingly, the *mal*P gene of *P. torridus* is clustered with a probable α-mannosidase gene (ORF 984) and this gene order could not be found in any of the organisms that possess homologs for one of two ORFs. The two genes were found on the same DNA strand and were separated by 40 bp. This, together with their functional relatedness suggests that they are cotranscribed.

Amino acid sequence alignment of MalP with representatives of family 31 showed the presence of the conserved active site pattern [GF]-[LIVMF]-W-x-D-M-[NSA]-E (PROSITE database). The aspartic acid residue has been implicated in the catalytic activity of several representatives of this family (Hermans *et al.*, 1991). A second signature pattern containing two conserved cysteine residues has been described for this enzyme group (PROSITE database). In both the *S. solfataricus and P. torridus* proteins this region could not be identified, which is in agreement with the phylogenetic position of these organisms as outliers in the family.

The partial biochemical characterisation of the recombinant MalP which was carried out in this work showed that the enzyme had temperature and pH optima very close to its homolog from *S. solfataricus*, i.e. 87 vs. 105°C, respectively, and pH 4.5 for both (Fig. 19; Haseltine *et al.*, 1997). At the optimal growth temperature of *P. torridus* MalP displayed less than 20 % of its maximal activity when pNP-α glucoside was used as the substrate. This is consistent with measurements of the native α-glucosidase activity in crude *P. torridus* cellular extracts, where similar values were obtained (data not shown). Thin layer chromatography analysis (TLC) of the reaction products showed that MalP had a preference for short maltosaccharides and confirmed its annotation as an α-glucosidase. In contrast to the *S. solfataricus* homolog however, MalP showed activity with starch (Fig. 20; Rolfsmeier *et al.*, 1995).

3.2. *P. torridus* glucose dehydrogenase (GdhA)

P. torridus glucose dehydrogenase catalyses the oxidation of glucose to gluconate via gluconolactone using NAD^+ and $NADP^+$ as cofactors:

$$\text{Glucose} + NAD(P)^+ \rightarrow \text{Gluconate} + NAD(P)H + H^+$$

Two types of glucose dehydrogenases have been described: pyrroloquinoline-quinone (PQQ)-dependent (EC 1.1.99.17) and pyridine-nucleotide-dependent glucose dehydrogenases (EC 1.1.1.47). PQQ-dependent enzymes have been found exclusively in gram-negative bacteria such as several species of *Gluconobacter, Acinetobacter, Klebsiella, Pseudomonas,* and *Escherichia* (Cleton-Jansen *et al.,* 1988, 1989). Pyridine-nucleotide-dependent glucose dehydrogenases have been isolated and characterized from all three domains of life: *Bacteria*, *Archaea*, and *Eukarya*. Pyridine-nucleotide-dependent enzymes typically display tetrameric structure and dual cosubstrate specifity for both NAD^+ and $NADP^+$ (Campbell *et al.*, 1982; Jany *et al.*, 1984; Giardina *et al.*, 1986, Smith *et al.*, 1989). Exceptions are the tetrameric enzymes of *Schizosaccharomyces pombe* (Tsai *et al.*, 1995), *Halobacterium halobium* (Sonawat *et al.*, 1990), and *Thermoplasma acidophilum* (Smith *et al.*, 1989) for which a high $NADP^+$ specificity has been described, the dimeric enzyme from *Haloferax mediterranei* (Bonete *et al.*, 1996), and the monomeric enzyme of *Corynebacterium* species (Kobayashi *et al.*, 1982). An interesting representative of this group is the glucose dehydrogenase of *Thermoproteus tenax* whose quaternary structure is dependent on the presence of the cosubstrate (Siebers *et al.*, 1997). According to primary structure analysis, PQQ-dependent and pyridine-nucleotide-dependent glucose dehydrogenases are not related (Cleton-Jansen *et al.*, 1988).

Based on amino acid sequence similarity, *P. torrridus* glucose dehydrogenase could be assigned as a member of the medium–chain alcohol/polyol dehydrogenase/reductase branch of the superfamily of pyridine-nucleotide-dependent alcohol/polyol/sugar dehydrogenases (Edwards *et al.*, 1996). The crystal structure of the glucose dehydrogenase from *T. acidophilum* has been reported and, due to the high degree of amino acid sequence similarity with the *P. torridus* ortholog (59.3 % identity), it is possible to make comparisons between the two enzymes at the structural level (John *et al.*, 1994; see section C.3.2.1).

3.2.1. Structural basis for stability

Surprisingly, it was observed that the purified enzyme was inactivated completely after incubation for 1h at conditions thought to be physiological for a cytoplasmic enzyme of *P. torridus* (60°C and pH 4.6). This finding suggested that some stabilizing factors could have been lost during the purification process. Indeed, the results obtained in this study indicate the critical importance of Zn^{2+} for the stability of GdhA. The resistance of GdhA against inactivation at high temperature as well as its stability at low pH were considerably increased in the presence of $ZnCl_2$, and this effect was abolished by the chelating agent EDTA. On the other hand, the presence of Zn^{2+} did not affect the activity of the enzyme, and even high concentrations of EDTA (20 mM) could not decrease the activity of GdhA in the standard assay. This is in contrast to the effect of EDTA on the glucose dehydrogenase from *Sulfolobus solfataricus*, where at 10 mM concentrations the reported decrease in activity is 60% (Lamble *et al.*, 2003). These observations could be explained by a very stable coordination of Zn^{2+} in the catalytic site and a more labile structural zinc in the *P. torridus* protein. This may also be the case for the glucose dehydrogenase from *T. acidophilum*. Based on the conservation of the zinc–binding sequence of both enzymes (Fig. 20), including the cysteine and aspartate residues involved in coordination of the metal ions, the structural basis of zinc binding in *P. torridus* GdhA is probably similar to the situation found in *T. acidophilum* glucose dehydrogenase. John *et al.* (1994) have shown that in the *T. acidophilum* glucose dehydrogenase the catalytic and nucleotide-binding domains are separated by a deep active site cleft, the putative catalytic zinc being at the bottom of the cleft and a lobe containing the structural zinc at the mouth of the cleft and thus exposed to the solvent (John *et al.*, 1994; Fig. 36).

Similar observations have been reported for yeast alcohol dehydrogenase, another member of the medium-chain alcohol/polyol dehydrogenase family that carries structural similarities with *T. acidophilum* glucose dehydrogenase, where gradual chelating depleted the enzyme first of the structural and then of the catalytic zinc (Magonet *et al.*, 1992).

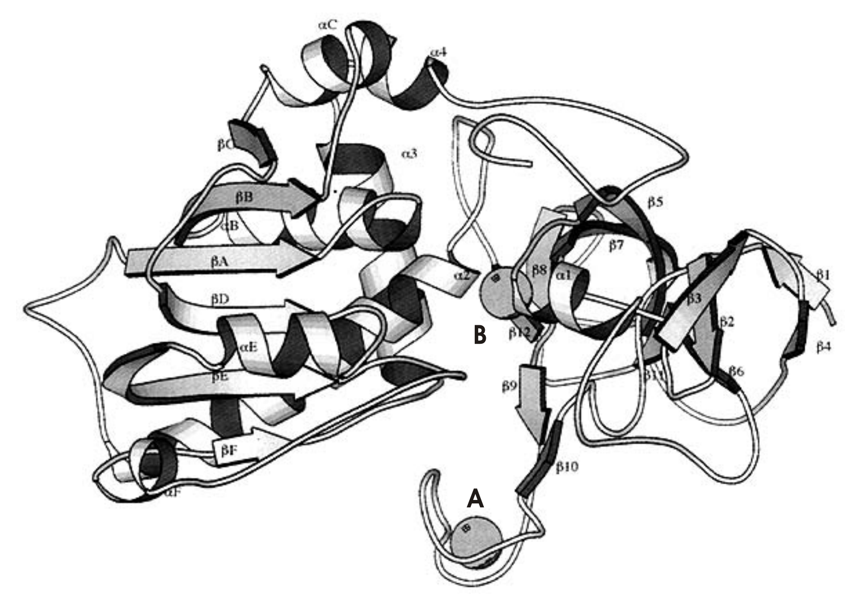


Image from John et al.,Structure 1994, 2:385-393.

Fig. 36. Schematic representation of the *T. acidophilum* glucose dehydrogenase monomer (John et al., 1994). The two zinc ions are shown as spheres.
A: structural zinc. B: catalytic zinc.

Active *P. torridus* GdhA has a tetrameric quaternary structure which is found in most members of this family of glucose dehydrogenases (section C.3.2.2). It has been argued previously that the role of the structural zinc is to stabilise the quaternary structure of *T. acidophilum* glucose dehydrogenase (John *et al.*, 1994). However, no change in the quaternary structure of *P. torridus* GdhA destabilized by EDTA treatment was observed (data not shown), indicating that in *P. torridus* GdhA the structural zinc is only responsible for stabilising the tertiary structure of the enzyme.

3.2.2. Native glucose/galactose dehydrogenase activity in P. torridus

The glucose (galactose) dehydrogenase activity measured in *P. torridus* crude cellular extracts turned out to have very similar characteristics with the recombinant protein, i.e. pH and temperature optima, $NADP^+$ **:** NAD^+ and glucose **:** galactose activity ratios (Table 15). Furthermore, after zymogram staining of proteins separated on a native PAGE gel for glucose dehydrogenase activity, the purified recombinant enzyme was undistinguishable from the band of the *P. torridus* crude extract (Fig. 29). Thus it was assumed that the GdhA protein indeed represents the prominent glucose dehydrogenase activity in *P. torridus* cells under the growth conditions employed in this study. Considering the presence of two additional putative glucose dehydrogenase ORFs in the *P. torridus* genome, however, further experiments are needed to unequivocally show the expression of *gdhA* in *P. torridus*.

4. Physiological role of GdhA and the Entner-Doudoroff pathway

The availability of annotated complete genome sequences from various organisms and its complementation with biochemical evidence permits a more detailed view and comparison to be made of the highly plastic variants of glucose degradation in different species. Glucose dehydrogenase is the first enzyme in a variant of the Entner-Doudoroff pathway, involving non-phosphorylated intermediates. The functionality of the non-phosphorylated variant of the Entner–Doudoroff pathway (ED) has been shown in the thermoacidophilic Archaea *Sulfolobus solfataricus* (De Rosa *et al.*, 1984), *Sulfolobus acidocaldarius* (Selig *et al.*, 1997), *Thermoplasma acidophilum* (Budgen *et al.*, 1986), as well as, with some modifications, in *Thermoproteus tenax* (Siebers *et al.*, 1997). Genome-based metabolic pathway reconstruction together with the identification of a native glucose dehydrogenase activity in cell free extracts has suggested its presence also in *P. torridus* (section C.1.3.9). In this pathway, phosphorylation takes place only at the level of glycerate, in contrast to the semi-phosphorylated ED, described to operate in halophilic Archaea, where phosphorylation occurs at the level of 2-keto-3-deoxy gluconate (KDG) (Johnsen *et al.*, 2001; Fig. 37).

It has been proposed that the utilisation of the different variants of the ED pathway is connected with the adaptation to high temperatures due to the instability of several key intermediates of the EMP. The reported half-lives at 60°C for glyceraldehyde-3-phosphate (GAP), dihydroxyacetone phosphate (DHAP) and 1,3-diphosphoglycerate are 14.5, 79.4 and 1.6 minutes, respectively (Dörr *et al*., 2003). The most heat-labile compound, 1,3-diphosphoglycerate, can be avoided by directly converting GAP to 3-phosphoglycerate with a non-phosphorylating GAP dehydrogenase (*P. torridus* ORF 1226), while the use of the non-phosphorylated ED pathway circumvents also the formation of the other two heat-labile components (Fig. 37).

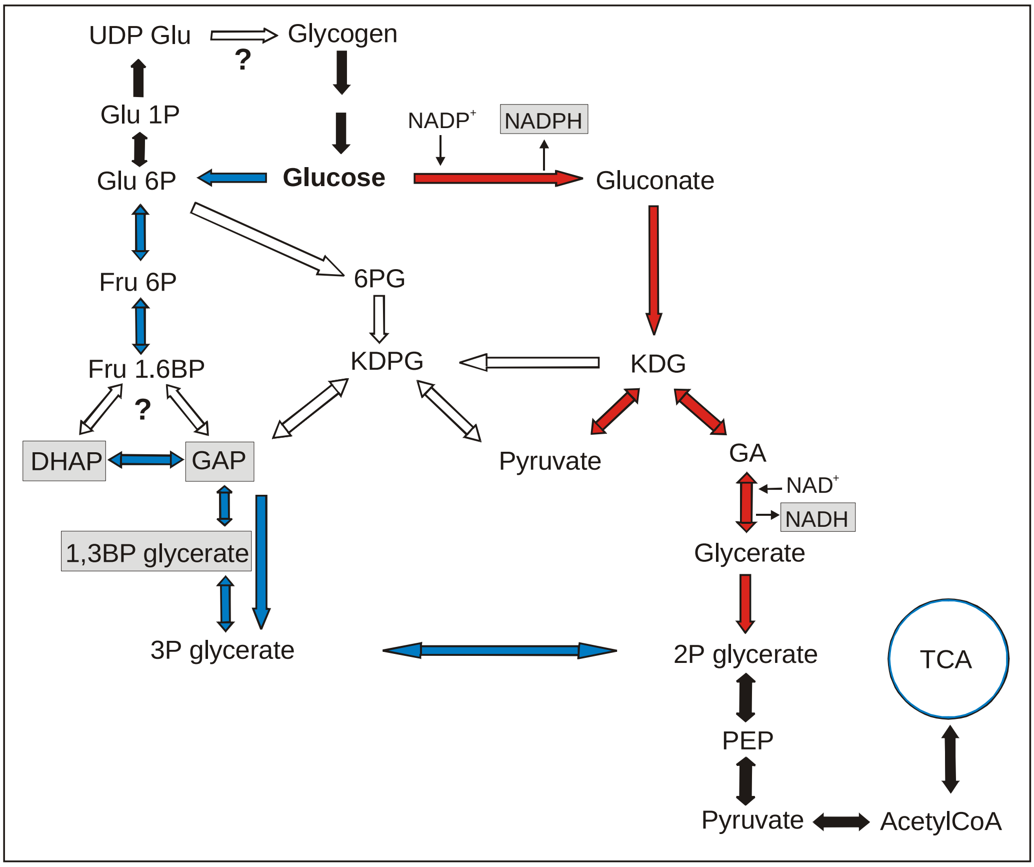


Fig. 37. The possible routes for glucose utilisation in *P. torridus* as inferred from genome data. The enzyme steps for which genes could be found are marked with filled arrows and the missing ones with empty arrows. In order to emphasise the variable possibilities of glucose degradation that are known, the classical and the semi-phosphorylated ED pathways are shown in addition. The intermediates known to be unstable at high temperature (60°C) are boxed.

However, one of the products of the glucose dehydrogenase reaction, NADPH, is also known to be unstable at elevated temperatures and low pH values; other factors reported to affect its rate of degradation are the ionic strength and the presence of phosphate and acetate ions in the buffer system (Wu *et al.*, 1986). At the conditions considered to be physiological for *P. torridus* (pH 4.5 and 60°C), NADPH showed dramatically decreased stability, the most important factor being the hydronium ion concentration (Fig. 27). The half-life of NADPH determined at 60°C was 1.9 min at pH 4.6 and 49 min at pH 6.5 (section C.3.2.3.6). Although to a lesser extent, NADH has also been reported to be unstable at elevated temperatures (Wu *et al.*, 1986). Therefore, at least in *P. torridus*, the enzymes consuming NAD(P)H are expected to have a high affinity for these cofactors.

There are several lines of evidence that support the hypothesis that glucose degradation in *P. torridus* might not be restricted only to the ED pathway and that, at least for gluconeogenesis, the EMP route might be utilised:

1. Neither *P. torridus* nor *T. acidophilum* appear to contain a classical glycogen synthase homolog and experimental data is missing as to whether *P. torridus* is able to synthesize storage polysaccharides. However, most biosynthetic genes were identified in the genome sequence, i.e. those coding for a phosphoglucomutase and a glucose-1-phosphate uridylyltransferase. Therefore, considering the large reservoir of genes with unknown function, i.e. 553 ORFs, it is plausible that another protein has evolved to catalyze the missing reaction. This process has been referred to as non-orthologous gene displacement (Koonin *et al.*, 1996). Also, genes apparently involved in hydrolytic glycogen degradation could be identified - a putative glycogen debranching enzyme gene was found in a probable operon with a family 57 α-amylase and a phosphofructokinase gene adjacent to the cluster (Fig. 32). It can be concluded therefore that *P. torridus* is probably able to synthesize storage compounds despite the lack of a classical glycogen synthase gene.

The results from testing the substrate specificity of the purified recombinant GdhA indicate a relatively strict range of substrates for this enzyme. Nevertheless, GdhA was considerably active with D-galactose, and it displayed approximately twofold higher affinity to this substrate (K_m= 4.5 mM) compared to D-glucose. The discovery of a `promiscuous´ Entner–Doudoroff pathway in *S. solfataricus* by Lamble *et al.* (2003) suggested that in this organism the utilisation of glucose and galactose is carried out by the same enzymes, which lack facial selectivity. Based on the observed activity of GdhA with galactose, such a promiscuity cannot be excluded also in *P. torridus*. Therefore, assuming that the *P. torridus* KDG aldolase also does not discriminate between the two stereoisomers, the use of this pathway for gluconeogenesis would result in a mixture of glucose and galactose.

At the intracellular pH and temperature thought to be physiological for *P. torridus* the equilibrium between gluconate/galactonate and their lactone forms (gamma- and delta-) is strongly shifted towards the lactones. This would require either the presence of a lactonase or a high activity of the next enzyme in the pathway – gluconate/galactonate dehydratase. For both these enzymes probable genes could be identified in the genome. Moreover, when the purified GdhA was used, it was not possible to measure *in vitro* the reverse reaction, i.e. the formation of glucose from gluconate or gluconolactone (not shown).

2. In order to assume a gluconeogenetic role of the EMP in *P. torridus*, alternative enzymes have to be present for at least two virtually irreversible steps of this pathway, i.e. for the phosphofructokinase (PFK) and for the pyruvate kinase (PYK) reactions. The two enzymes that catalyse these reactions in the opposite direction and are thus characteristic for gluconeogenesis are the fructose-1,6-bisphosphatase (FBP) and the phosphoenolpyruvate synthase (PPS) which are found in most arhaeal genomes (Verhees *et al.*, 2003). The *P. torridus* ORF 158 displays significant amino acid sequence similarity (64 % identity) to the recently described novel type V FBP from *Thermococcus kodakaraensis* (Rashid *et al.*, 2002). The enzyme was shown to catalyse the irreversible dephosphorylation of fructose-1,6-bisphosphate to fructose-6-phosphate. The other irreversible EMP step that has to be overcome in the gluconeogenetic direction is the phosphorylation of pyruvate to phosphoenolpyruvate (PEP). A homolog of the classical PPS sequence could be found in the *P. torridus* genome (ORF 876). Despite the ongoing debate about the true function of PPS in Archaea, it is nevertheless considered to be a gluconeogenetic enzyme (Schut *et al.*, 2003).

3. Finally, maintaining the coding capacity for both EMP and ED pathways in a genome that is subjected to selective pressure favouring small genome size reflects their importance. It also suggests that their presence may not only be metabolic parallelism but rather a mechanism that confers advantage in a hot and acidic environment characterised by an irregular supply of energy sources.

When the energy yield obtained from each of the alternative pathways is compared it can be concluded that the processing of glucose to pyruvate via the EMP is the most profitable: while the net ATP gain is 2 mol/mol of glucose in the EMP, the utilisation of glucose via the classical ED generates 1 mol of ATP/mol of glucose, and in the non-phosphorylated ED no ATP is gained, at least at the substrate level. Assuming that in *P. torridus* both the EMP and the non-phosphorylated ED are present, it would be interesting to investigate whether the extent of utilisation of the possible pathways is connected with the growth temperature.

Very little is known about the regulation at the protein and gene levels of the ED and EMP pathways in archaea. The data obtained in the current work suggests that *P. torridus* GdhA has a little allosteric capacity (section C.3.2.3.1). The observed 30% decrease in activity by 5 mM ATP is of questionable physiological importance due to the high effective concentration of ATP necessary to observe this effect.

In summary, the results from this work demonstrate that the complete genome sequence of *P. torridus* can be used as an informational resource for studying thermoacidophilic adaptation. The data obtained from the genome is a crucial prerequisite for further investigation of the extreme acidophile *P. torridus*, i.e. analysis of the structure and function of selected enzymes, of the organism’s metabolism, and of the regulation of gene expression.

E. Summary

- In the frame of the current work the complete genome sequence of the archaeon *P. torridus* was determined. Sequence assembly, editing, gap closure and annotation were done in close collaboration with Dr. O. Fütterer (Department of Applied Microbiology, University of Göttingen) and the Göttingen Genomics Laboratory, with further input at the annotation stage from collaboration partners at the Department of Technical Microbiology (Prof. G. Antranikian) of the Technical University of Hamburg.
- The organism was found to contain a single chromosome of 1,545,900 base pairs and thus represents the smallest genome of a non-parasitic heterotrophic organism sequenced do far. After assembly, the average sequence redundancy of the genome was 9.4 fold and the probability for an error below 1 in 2,000,000.
- In the genome, 1,535 ORFs could be identified, and for 64 % of them a probable function could be assigned. Metabolic pathway reconstruction allowed ordering of these genes in the main functional categories – protein synthesis, transport, energy metabolism, nucleotide, DNA and RNA metabolism. A characteristic feature observed in this genome is the high coding density of 91.7 % which, together with the small genome size, is probably a result of the selective pressure exerted by the extreme environment.
- Scouring the genome for genes for transporters revealed that 93 were secondary proton-driven transport systems, giving a ratio of secondary to primary transporters of 5.6:1. This high ratio, compared to other microorganisms such as *Sulfolobus solfataricus* (2.7:1), *Escherichia coli* (2.6:1), *Pyrococcus horikoshii* (1.5:1) or *Thermotoga maritima* (0.5:1), is considered to be relevant to the adaptation of the organism to the acidic environment. On the other hand, nearly half of the ORFs necessary for the uptake of peptides, amino acids (34 ORFs) and sugars (32 ORFs) code for primary ATP-binding cassette (ABC)-transporter subunits, emphasising the necessity of high affinity transport for these substrates. Another transport system thought to be connected with the extreme lifestyle of *P. torridus* is the K^+ translocating ATPase whose physiological role is probably to maintain a $\Delta\psi$ that is positive inside and in this way to protect the cell against the steep proton gradient across the cytoplasmic membrane.
- An important conclusion obtained from investigating the distribution of *P. torridus* ORFs' homologs is that phylogenetically distant groups of archaea, isolated

from similar habitats, share an unexpectedly large pool of genes. Thus, the thermoacidophiles form a distinct group where ecological closeness in extremely acidic and hot habitats has had a pronounced effect on the evolutionary “shaping” of their genomes. For the occurrence of many of the common genes of thermoacidophiles, genetic exchange rather than common descent must have played an important role.

- Many of the traits thought to be relevant to the survival of *P. torridus* in conditions of high acidity seem to have been acquired by lateral gene transfer from distantly related taxa. These include key enzymes from the organic acid degradation pathways, the main components of the electron transport chain and genes connected with protection against oxidative damage.
- In order to further investigate the adaptation mechanisms allowing the organism to survive at low pH values, several *P. torridus* ORFs were selected for heterologous expression. Standard (*E. coli*) as well as alternative (*S. solfataricus* and *S. cerevisiae*) expression systems were used. The most important factor that affected the proper folding of the *P. torridus* polypeptides in *E. coli* was determined to be the rate of their synthesis. Two recombinant enzymes were produced and purified, i.e., α-glucosidase (MalP) and glucose-1-dehydrogenase (GdhA).
- Biochemical characterisation of the recombinant GdhA showed that the active enzyme is a tetramer, does not discriminate between glucose and galactose as substrates, uses $NADP^+$ as cofactor and requires Zn^{2+} in order to be stable under physiological conditions. Furthermore, GdhA showed considerable stability in the presence of organic solvents: overnight incubation (14h) at room temperature with 50% v/v of acetone, methanol or ethanol did not result in a detectable loss of activity.
- The data obtained from studying the recombinant GdhA together with the information extracted from the genome sequence suggests that, in *P. torridus*, the non-phosphorylated Entner-Doudoroff pathway is responsible for the degradation of glucose, while the enzymes of the Embden-Meyerhof-Parnas pathway are most probably involved in gluconeogenesis.
- The *P. torridus* α-glucosidase enzyme MalP was found to be active over a broad pH and temperature range. Maximal activity in the standard 10 min assay was measured at 87°C, well above the growth temperature optimum of *P.torridus*. Proteins with a detectable degree of amino acid sequence similarity to MalP could be found only in distantly related taxa (from crenarchaea, bacteria and eukaryotes).

F. References

Albers SV, Van de Vossenberg JL, Driessen AJ, Konings WN. Bioenergetics and solute uptake under extreme conditions. Extremophiles. 2001 Oct; **5** (5):285-94.

Altschul SF, Lipman DJ. Protein database searches for multiple alignments. Proc Natl Acad Sci U S A. 1990 Jul; **87** (14)**:**5509-13.

Amann, E., Brosius, J. & Ptashne, M. Vectors bearing a hybrid *trp-lac* promoter useful for regulated expression of cloned genes in *Escherichia coli*. Gene 1983; **25** (2-3), 167-78.

Angelaccio S, Chiaraluce R, Consalvi V, Buchenau B, Giangiacomo L, Bossa F, Contestabile R. Catalytic and thermodynamic properties of tetrahydromethanopterin-dependent serine hydroxymethyltransferase from *Methanococcus jannaschii*. J Biol Chem. 2003 Oct 24; **278** (43):41789-97. Epub 2003 Aug 05.

Badger JH, Olsen GJ. CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol. 1999 Apr; **16** (4)**:**512-24.

Baik SH, Ide T, Yoshida H, Kagami O, Harayama S. Significantly enhanced stability of glucose dehydrogenase by directed evolution. Appl Microbiol Biotechnol. 2003 May; **61** (4):329-35. Epub 2003 Mar 05.

Batas B, Chaudhuri JB. Considerations of sample application and elution during size-exclusion chromatography-based protein refolding. J Chromatogr A. 1999 Dec 24; **864** (2):229-36.

Bates, E.J.; Heaton, G.M.; Taylor, C.; Kernohan, J.C.; Cohen, P. Debranching enzyme from rabbit skeletal muscle: evidence for the location of two active centres on a single polypeptide chain. FEBS Lett. 1975; **58**:181-185.

Bearson S, Bearson B, Foster JW. Acid stress responses in enterobacteria. FEMS Microbiol Lett. 1997 Feb 15; **147** (2):173-80.

Bitan-Banin G, Ortenberg R, Mevarech M. Development of a gene knockout system for the halophilic archaeon *Haloferax volcanii* by use of the pyrE gene. J Bacteriol. 2003 Feb; **185** (3)**:**772-8.

Bonete MJ, Pire C, LLorca FI, Camacho ML. Glucose dehydrogenase from the halophilic Archaeon *Haloferax mediterranei*: enzyme purification, characterisation and N-terminal sequence. FEBS Lett. 1996 Apr 1; **383** (3):227-9.

Bonfield, J.K. and Staden, R. The application of numerical estimates of base calling accuracy to DNA sequencing projects. Nucleic Acids Res. 1995, **23**, 1406-1410.

Boucher Y, Douady CJ, Papke RT, Walsh DA, Boudreau ME, Nesbo CL, Case RJ, Doolittle WF. Lateral gene transfer and the origins of prokaryotic groups. Annu Rev Genet. 2003; **37:**283-328.

Bradford, M. A rapid and sensitive method for quantification of microgram quantities of protein utilizing the principle of protein-dye binding. Anal Biochem 1976, **72**, 248-254.

Brock, T.D. Thermophilic Microorganisms and Life at High Temperatures. 1978 (Springer, Berlin,) 255-302.

Brosius, J., Ullrich, A., Raker, M.A., Gray, A., Dull, T.J., Gutell, R.R. & Noller, H.F. Construction and fine mapping of recombinant plasmids containing the rrnB ribosomal RNA operon of *E. coli*. Plasmid 1981 **6** (1),112-8.

Budgen N and Danson M. Metabolism of glucose via a modified Entner-Doudoroff pathway in the thermoacidophilic archaebacterium *Thermoplasma acidophilum*. FEBS 1986, **196-2**, 207.

Bullock, W. O., Fernandez, J. M. & Short, J. M. XL1-Blue: a high efficiency plasmid DNA transforming *rec*A *Escherichia coli* strain with beta-galactosidase selection. BioTechniques 1987**, 5**: 376-379.

Campbell DP, Carper WR, Thompson RE. Bovine liver glucose dehydrogenase: isolation and characterization. Arch Biochem Biophys. 1982 Apr 15; **215** (1):289-301.

Chaudhuri JB. Refolding recombinant proteins: process strategies and novel approaches. Ann N Y Acad Sci. 1994 May 2; **721** :374-85.

Ciaramella M, Pisani FM, Rossi M. Molecular biology of extremophiles: recent progress on the hyperthermophilic archaeon *Sulfolobus*. Antonie Van Leeuwenhoek. 2002 Aug; **81** (1-4)**:**85-97.

Cleton-Jansen AM, Goosen N, Wenzel TJ, van de Putte P. Cloning of the gene encoding quinoprotein glucose dehydrogenase from A*cinetobacter calcoaceticus*: evidence for the presence of a second enzyme. J Bacteriol. 1988 May; **170** (5):2121-5.

Constantino, H.R.; Brown, S.H.; Kelly, R.M. Purification and characterization of an alpha-glucosidase from a hyperthermophilic archaebacterium, *Pyrococcus furiosus*, exhibiting a temperature optimum of 105 to 115°C J. Bacteriol. 1990;**172:**3654-3660

Cryer, D. R., Eccleshall, R. & Marmur, J. Isolation of yeast DNA. Methods Cell Biol. 1975 **12**, 39-44.

Darland G, Brock TD, Samsonoff W, Conti SF. A thermophilic, acidophilic mycoplasma isolated from a coal refuse pile. Science. 1970 Dec 25; **170** (965):1416-8.

D'Auria S, Di Cesare N, Gryczynski Z, Gryczynski I, Rossi M, Lakowicz JR. A thermophilic apoglucose dehydrogenase as nonconsuming glucose sensor. Biochem Biophys Res Commun. 2000 Aug 11; **274** (3):727-31.

De Rosa M, Gambacorta A, Nicolaus B, Giardina P, Poerio E, Buonocore V. Glucose metabolism in the extreme thermoacidophilic archaebacterium *Sulfolobus solfataricus*. Biochem J. 1984 Dec 1; **224** (2):407-14.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. Improved microbial gene identification with GLIMMER. Nucleic Acids Res. 1999 Dec 1; **27** (23):4636-41.

DeLong EF. Extreme genomes. Genome Biol. 2000; **1** (6):REVIEWS1029. Epub 2000 Dec 04.

Deppenmeier U, Johann A, Hartsch T, Merkl R, Schmitz RA, Martinez-Arias R, Henne A, Wiezer A, Baumer S, Jacobi C, Bruggemann H, Lienard T, Christmann A, Bomeke M, Steckel S, Bhattacharyya A, Lykidis A, Overbeek R, Klenk HP, Gunsalus RP, Fritz HJ, Gottschalk G. The genome of *Methanosarcina mazei*: evidence for lateral gene transfer between bacteria and archaea. J Mol Microbiol Biotechnol. 2002 Jul; **4** (4):453-61.

Dörr C, Zaparty M, Tjaden B, Brinkmann H, Siebers B. The hexokinase of the hyperthermophile *Thermoproteus tenax.* ATP-dependent hexokinases and ADP-dependent glucokinases, teo alternatives for glucose phosphorylation in Archaea. J Biol Chem. 2003 May 23; **278** (21):18744-53. Epub 2003 Mar 07.

Edwards, KJ., Barton, JD., Rossjohn, J., Thorn, JM., Taylor, GL. & Ollis, DL. () Structural and sequence comparisons of quinone oxidoreductase, zeta-crystallin, and glucose and alcohol dehydrogenases. *Arch Biochem Biophys*. 1996; **328:** 173-183.

Elble R. A simple and efficient procedure for transformation of yeasts. Biotechniques. 1992 Jul; **13** (1):18-20.

Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. 1998 Mar; **8** (3):186-94.

F**leischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, et al.** Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. 1995 Jul 28; **269** (5223):496-512.

Foster JW. When protons attack: microbial strategies of acid adaptation. Curr Opin Microbiol. 1999 Apr; **2** (2):170-4.

Frishman D, Mironov A, Mewes HW, Gelfand M. Combining diverse evidence for gene recognition in completely sequenced bacterial genomes. Nucleic Acids Res. 1998 Jun 15; **26** (12):2941-7.

Fujita Y, Ramaley R, Freese E. Location and properties of glucose dehydrogenase in sporulating cells and spores of *Bacillus subtilis.* J Bacteriol. 1977 Oct; **132** (1):282-93.

Gething MJ, Sambrook J. Protein folding in the cell. Nature. 1992 Jan 2; **355** (6355):33-45.

Giardina P, de Biasi MG, de Rosa M, Gambacorta A, Buonocore V. Glucose dehydrogenase from the thermoacidophilic archaebacterium *Sulfolobus solfataricus*. Biochem J. 1986 Nov 1; **239** (3):517-22.

Gibson TJ, Rosenthal A, Waterston RH. Lorist6, a cosmid vector with BamHI, NotI, ScaI and HindIII cloning sites and altered neomycin phosphotransferase gene expression. Gene. 1987; **53** (2-3):283-6.

Gribskov M, Burgess RR. Overexpression and purification of the sigma subunit of *Escherichia coli* RNA polymerase. Gene. 1983 Dec; **26** (2-3):109-18.

Harrison G. Expression of soluble heterologous proteins via fusion with NusA protein. Innovations. 2000 **11**: 4-7.

Harrowing SR, Chaudhuri JB. Effect of column dimensions and flow rates on size-exclusion refolding of beta-lactamase. J Biochem Biophys Methods. 2003 Jun 30; **56** (1-3):177-88.

Harrowing SR, Chaudhuri JB. Effect of column dimensions and flow rates on size-exclusion refolding of beta-lactamase. J Biochem Biophys Methods. 2003 Jun 30; **56** (1-3):177-88.

Hartl FU, Hayer-Hartl M. Molecular chaperones in the cytosol: from nascent chain to folded protein. Science. 2002 Mar 8; **295** (5561):1852-8.

Haseltine C, Montalvo-Rodriguez R, Carl A, Bini E, Blum P. Extragenic pleiotropic mutations that repress glycosyl hydrolase expression in the hyperthermophilic archaeon *Sulfolobus solfataricus*. Genetics. 1999 Aug; **152** (4):1353-61.

Hemmi H, Ikejiri S, Nakayama T, Nishino T. Fusion-type lycopene beta-cyclase from a thermoacidophilic archaeon *Sulfolobus solfataricus*. Biochem Biophys Res Commun. 2003 Jun 6; **305** (3):586-91.

Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. Biochem J. 1991 Dec 1; **280** (Pt 2):309-16.

Hermans M.M.P., Kroos M.A., van Beeumen J., Oostra B.A., Reuser A.J.J. Human lysosomal alpha-glucosidase. Characterization of the catalytic site. J. Biol. Chem. **266:**13507-13512(1991).

Hutchins AM, Holden JF, Adams MW. Phosphoenolpyruvate synthetase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol. 2001 Jan; **183** (2):709-15.

J**any KD, Ulmer W, Froschle M, Pfleiderer G.** Complete amino acid sequence of glucose dehydrogenase from *Bacillus megaterium*. FEBS Lett. 1984 Jan 2; **165**(1):6-10.

John J, Crennell SJ, Hough DW, Danson MJ, Taylor GL. The crystal structure of glucose dehydrogenase from *Thermoplasma acidophilum*. Structure. 1994 May 15; **2** (5):385-93.

Johnsen U, Selig M, Xavier KB, Santos H, Schonheit P. Different glycolytic pathways for glucose and fructose in the halophilic archaeon *Halococcus saccharolyticus*. Arch Microbiol. 2001 Jan; **175** (1):52-61.

Johnson , DB. Biodiversity and ecology of acidophilic microorganisms. FEMS Microbiology Ecology 1998, **27:** 307-317

Johnson DB, Okibe N, Roberto FF. Novel thermo-acidophilic bacteria isolated from geothermal sites in Yellowstone National Park: physiological and phylogenetic characteristics. Arch Microbiol. 2003 Jul; **180** (1):60-8. Epub 2003 Jun 07.

Jonuscheit M, Martusewitsch E, Stedman KM, Schleper C. A reporter gene system for the hyperthermophilic archaeon *Sulfolobus solfataricus* based on a selectable and integrative shuttle vector. Mol Microbiol. 2003 Jun; **48** (5):1241-52.

Kawarabayasi Y, Hino Y, Horikawa H, Jin-no K, Takahashi M, Sekine M, Baba S, Ankai A, Kosugi H, Hosoyama A, Fukui S, Nagai Y, Nishijima K, Otsuka R, Nakazawa H et al., Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7. DNA Res. 2001 Aug 31; **8** (4):123-40.

Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. Nature. 2004 Apr 8; **428** (6983):617-24. Epub 2004 Mar 07.

Kobayashi, Y.; Ueyama, H.; Horikoshi, K. Comparative studies of NAD+-dependent maltose dehydrogenase and D-glucose dehydrogenase produced by two strains of alkalophilic *Corynebacterium species* Agric. Biol. Chem. 1982; **46**:2139-2142.

König H, Skorko R, Zillig W, Reiter WD. Glycogen in thermoacidophilic archaebacteria of the genera *Sulfolobus, Thermoproteus, Desulforococcus and Thermococcus*. Arch Microbiol. 1982. **132**: 297-303.

Kurtz S, Schleiermacher C. REPuter: fast computation of maximal repeats in complete genomes. Bioinformatics. 1999 May; **5:**426-7.

Lamble HJ, Heyer NI, Bull SD, Hough DW, Danson MJ. Metabolic pathway promiscuity in the archaeon *Sulfolobus solfataricus* revealed by studies on glucose dehydrogenase and 2-keto-3-deoxygluconate aldolase. J Biol Chem. 2003 Sep 5; **278** (36):34066-72. Epub 2003 Jun 24.

Lander ES, Waterman MS. Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics. 1988 Apr; **2** (3):231-9.

Lubben M, Warne A, Albracht SP, Saraste M. The purified SoxABCD quinol oxidase complex of *Sulfolobus acidocaldarius* contains a novel haem. Mol Microbiol. 1994 Jul; **13** (2):327-35.

Lucas S, Toffin L, Zivanovic Y, Charlier D, Moussard H, Forterre P, Prieur D, Erauso G. Construction of a shuttle vector for, and spheroplast transformation of, the hyperthermophilic archaeon *Pyrococcus abyssi*. Appl Environ Microbiol. 2002 Nov; **68** (11):5528-36.

Magonet E, Hayen P, Delforge D, Delaive E, Remacle J. Importance of the structural zinc atom for the stability of yeast alcohol dehydrogenase. Biochem J. 1992 Oct 15; **287** (Pt 2):361-5.

Makino Y, Negoro S, Urabe I, Okada H. Stability-increasing mutants of glucose dehydrogenase from *Bacillus megaterium* IWG3. J Biol Chem. 1989 Apr 15; **264** (11):6381-5.

Mitchell P. Chemiosmotic coupling in oxidative and photosynthetic phosphorylation. Biol Rev Camb Philos Soc. 1966 Aug; **41** (3):445-502.

Nishihara K, Kanemori M, Kitagawa M, Yanagi H, Yura T. Chaperone coexpression plasmids: differential and synergistic roles of DnaK-DnaJ-GrpE and GroEL-GroES in assisting folding of an allergen of Japanese cedar pollen, Cryj2, in *Escherichia coli.* Appl Environ Microbiol. 1998 May; **64** (5):1694-9.

Norris P.R., Johnson D.B. Acidophilic microorganisms. In: Extremophiles: Microbial life in extreme environments. 1998 Wiley, New York, NY.

Overbeek R, Larsen N, Pusch GD, D'Souza M, Selkov E Jr, Kyrpides N, Fonstein M, Maltsev N, Selkov E. WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. Nucleic Acids Res. 2000 Jan 1; **28** (1):123-5.

Palm P, Schleper C, Grampp B, Yeats S, McWilliam P, Reiter WD, Zillig W. Complete nucleotide sequence of the virus SSV1 of the archaebacterium *Sulfolobus shibatae*. Virology. 1991 Nov; **185** (1):242-50.

Piller K, Daniel RM, Petach HH. Properties and stabilization of an extracellular alpha-glucosidase from the extremely thermophilic archaebacteria *Thermococcus* strain AN1: enzyme activity at 130 degrees C. Biochim Biophys Acta. 1996 Jan 4; **1292** (1):197-205.

Raasch C, Streit W, Schanzer J, Bibel M, Gosslar U, Liebl W. *Thermotoga maritima* AglA, an extremely thermostable NAD+-, Mn2+-, and thiol-dependent alpha-glucosidase. Extremophiles. 2000 Aug; **4** (4):189-200.

Rashid N, Imanaka H, Kanai T, Fukui T, Atomi H, Imanaka T. A novel candidate for the true fructose-1,6-bisphosphatase in archaea. J Biol Chem. 2002 Aug 23; **277** (34):30649-55. Epub 2002 Jun 13.

Reeve JN. Archaebacteria then ... Archaes now (are there really no archaeal pathogens?). J Bacteriol. 1999 Jun; **181** (12):3613-7.

Rodionov DA, Vitreschak AG, Mironov AA, Gelfand MS. Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. J Biol Chem. 2003 Oct 17; **278** (42):41148-59. Epub 2003 Jul 17.

Rolfsmeier M, Blum P. Purification and characterization of a maltase from the extremely thermophilic crenarchaeote *Sulfolobus solfataricus*. J Bacteriol. 1995 Jan; **177** (2):482-5.

Ruepp A, Graml W, Santos-Martinez ML, Koretke KK, Volker C, Mewes HW, Frishman D, Stocker S, Lupas AN, Baumeister W. The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*. Nature. 2000 Sep 28; **407** (6803):508-13.

Sambrook J and Russel D. Molecular Cloning: A Laboratory Manual*(Third Edition)* 2001, Cold Spring Harbor Laboratory Press.

Sambrook, J., Fritsch, E. F. & Maniatis, T. *Molecular Cloning: a Laboratory Manual*, 2nd edn. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory**,** 1989.

Schäfer G, Engelhard M, Muller V. Bioenergetics of the Archaea. Microbiol Mol Biol Rev. 1999 Sep; **63** (3):570-620.

Schafer K, Magnusson U, Scheffel F, Schiefner A, Sandgren MO, Diederichs K, Welte W, Hulsmann A, Schneider E, Mowbray SL. X-ray structures of the maltose-maltodextrin-binding protein of the thermoacidophilic bacterium *Alicyclobacillus acidocaldarius* provide insight into acid stability of proteins. J Mol Biol. 2004 Jan 2; **335** (1):261-74.

Schleif R. Regulation of the L-arabinose operon of *Escherichia coli*. Trends Genet. 2000 Dec; **16** (12):559-65.

Schleper C, Kubo K, Zillig W. The particle SSV1 from the extremely thermophilic archaeon *Sulfolobus* is a virus: demonstration of infectivity and of transfection with viral DNA. Proc Natl Acad Sci U S A. 1992 Aug 15; **89** (16):7645-9.

Schleper C, Puehler G, Holz I, Gambacorta A, Janekovic D, Santarius U, Klenk HP, Zillig W. *Picrophilus* gen. nov., fam. nov.: a novel aerobic, heterotrophic, thermoacidophilic genus and family comprising archaea capable of growth around pH 0. J Bacteriol. 1995 Dec; **177** (24):7050-9.

Schut GJ, Brehm SD, Datta S, Adams MW. Whole-genome DNA microarray analysis of a hyperthermophile and an archaeon: *Pyrococcus furiosu*s grown on carbohydrates or peptides. J Bacteriol. 2003 Jul; **185** (13):3935-47.

Scott AI, Roessner CA. Biosynthesis of cobalamin (vitamin B(12)). Biochem Soc Trans. 2002 Aug; **30** (4):613-20.

Selig M, Xavier KB, Santos H, Schonheit P. Comparative analysis of Embden-Meyerhof and Entner-Doudoroff glycolytic pathways in hyperthermophilic archaea and the bacterium *Thermotoga*. Arch Microbiol. 1997 Apr; **167** (4):217-32.

Selkov E, Overbeek R, Kogan Y, Chu L, Vonstein V, Holmes D, Silver S, Haselkorn R, Fonstein M. Functional analysis of gapped microbial genomes: amino acid metabolism of *Thiobacillus ferrooxidans*. Proc Natl Acad Sci U S A. 2000 Mar 28; **97** (7):3509-14.

Serour E, Antranikian G. Novel thermoactive glucoamylases from the thermoacidophilic Archaea *Thermoplasma acidophilum, Picrophilus torridus and Picrophilus oshimae*. Antonie Van Leeuwenhoek. 2002 Aug; **81** (1-4):73-83.

She Q, Singh RK, Confalonieri F, Zivanovic Y et al., The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc Natl Acad Sci U S A. 2001 Jul 3; **98** (14):7835-40. Epub 2001 Jun 26.

Siebers B, Tjaden B, Michalke K, Dorr C, Ahmed H, Zaparty M, Gordon P, Sensen CW, Zibat A, Klenk HP, Schuster SC, Hensel R. Reconstruction of the central carbohydrate metabolism of *Thermoproteus tenax* by use of genomic and biochemical data. J Bacteriol. 2004 Apr; **186** (7):2179-94.

Siebers B, Wendisch VF, Hensel R. Carbohydrate metabolism in *Thermoproteus tenax*: in vivo utilization of the non-phosphorylative Entner-Doudoroff pathway and characterization of its first enzyme, glucose dehydrogenase. Arch Microbiol. 1997 Aug; **168** (2):120-7.

Smith LD, Budgen N, Bungard SJ, Danson MJ, Hough DW. Purification and characterization of glucose dehydrogenase from the thermoacidophilic archaebacterium *Thermoplasma acidophilum*. Biochem J. 1989 Aug 1; **261** (3):973-7.

Sonawat HM, Srivastava S, Swaminathan S, Govil G. Glycolysis and Entner-Doudoroff pathways in *Halobacterium halobium*: some new observations based on 13C NMR spectroscopy. Biochem Biophys Res Commun. 1990 Nov 30; **173** (1):358-62.

Speelmans G, Poolman B, Abee T, Konings WN. Energy transduction in the thermophilic anaerobic bacterium *Clostridium fervidus* is exclusively coupled to sodium ions. Proc Natl Acad Sci U S A. 1993 Sep 1; **90** (17):7975-9.

Spudich JL, Yang CS, Jung KH, Spudich EN. Retinylidene proteins: structures and functions from archaea to humans. Annu Rev Cell Dev Biol. 2000; **16** :365-92.

Staden R, Beal KF, Bonfield JK. The Staden package, 1998. Methods Mol Biol. 2000; **132:**115-30.

Stedman KM, Schleper C, Rumpf E, Zillig W. Genetic requirements for the function of the archaeal virus SSV1 in *Sulfolobus solfataricus*: construction and testing of viral shuttle vectors. Genetics. 1999 Aug; **152** (4):1397-405.

Tech M, Merkl R. YACOP: Enhanced gene prediction obtained by a combination of existing methods. In Silico Biol. 2003; **3** (4):441-51.

Tersteegen A, Linder D, Thauer RK, Hedderich R. Structures and functions of four anabolic 2-oxoacid oxidoreductases in *Methanobacterium thermoautotrophicum.* Eur J Biochem. 1997 Mar 15; **244** (3):862-8.

Tettelin H, Radune D, Kasif S, Khouri H, Salzberg SL. Optimized multiplex PCR: efficiently closing a whole-genome shotgun sequencing project. Genomics. 1999 Dec 15; **62** (3):500-7.

Thompson RE, Carper WR. Glucose dehydrogenase from pig liver. I. Isolation and purification. Biochim Biophys Acta. 1970 Mar 18;**198** (3):397-406.

Tomb JF, White O, Kerlavage AR, Clayton RA, Sutton GG, Fleischmann RD, Ketchum KA, Klenk HP, Venter JC, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature. 1997 Aug 7; **388** (6642):539-47.

Tsai CS, Ye HG, Shi JL. Carbon-13 NMR studies and purification of gluconate pathway enzymes from *Schizosaccharomyces pombe*. Arch Biochem Biophys. 1995 Jan 10; **316** (1):155-62.

Tucker DL, Tucker N, Conway T. Gene expression profiling of the pH response in *Escherichia coli*. J Bacteriol. 2002 Dec; **184** (23):6551-8.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM, Solovyev VV, Rubin EM, Rokhsar DS, Banfield JF. Community structure and metabolism through reconstruction of microbial genomes from the environment. Nature. 2004 Mar 4; **428** (6978):37-43. Epub 2004 Feb 01.

van de Vossenberg, Driessen AJ, Zillig W, Konings WN. Bioenergetics and cytoplasmic membrane stability of the extremely acidophilic, thermophilic archaeon *Picrophilus oshimae*. Extremophiles. 1998 May; **2** (2):67-74.

Venter JC. E. coli sequencing. Science. 1995 Feb 3;**267(5198):**601.

Verhees CH, Kengen SW, Tuininga JE, Schut GJ, Adams MW, De Vos WM, Van Der Oost J. The unique features of glycolytic pathways in Archaea. Biochem J. 2003 Oct 15; **375** (Pt 2):231-46.

Vuillard L, Rabilloud T, Goldberg ME. Interactions of non-detergent sulfobetaines with early folding intermediates facilitate in vitro protein renaturation. Eur J Biochem. 1998 Aug 15; **256** (1):128-35.

Wang RY, Kuo KC, Gehrke CW, Huang LH, Ehrlich M. Heat- and alkali-induced deamination of 5-methylcytosine and cytosine residues in DNA. Biochim Biophys Acta. 1982 Jun 30; **697** (3):371-7.

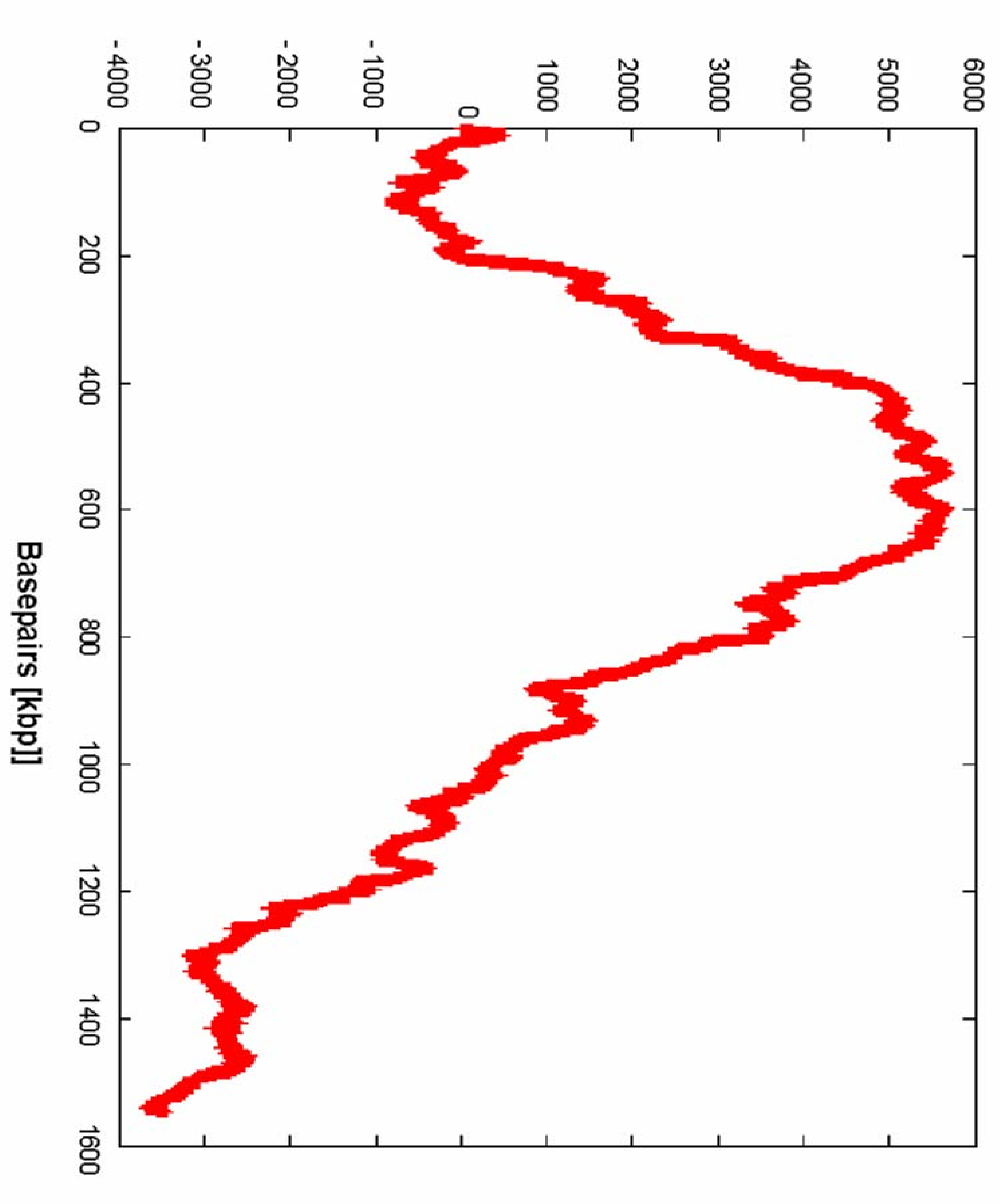
Woese CR, Kandler O, Wheelis ML. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A. 1990 Jun; **87** (12):4576-9.

Wu JT, Wu LH, Knight JA. Stability of NADPH: effect of various factors on the kinetics of degradation. Clin Chem. 1986 Feb; **32** (2):314-9.

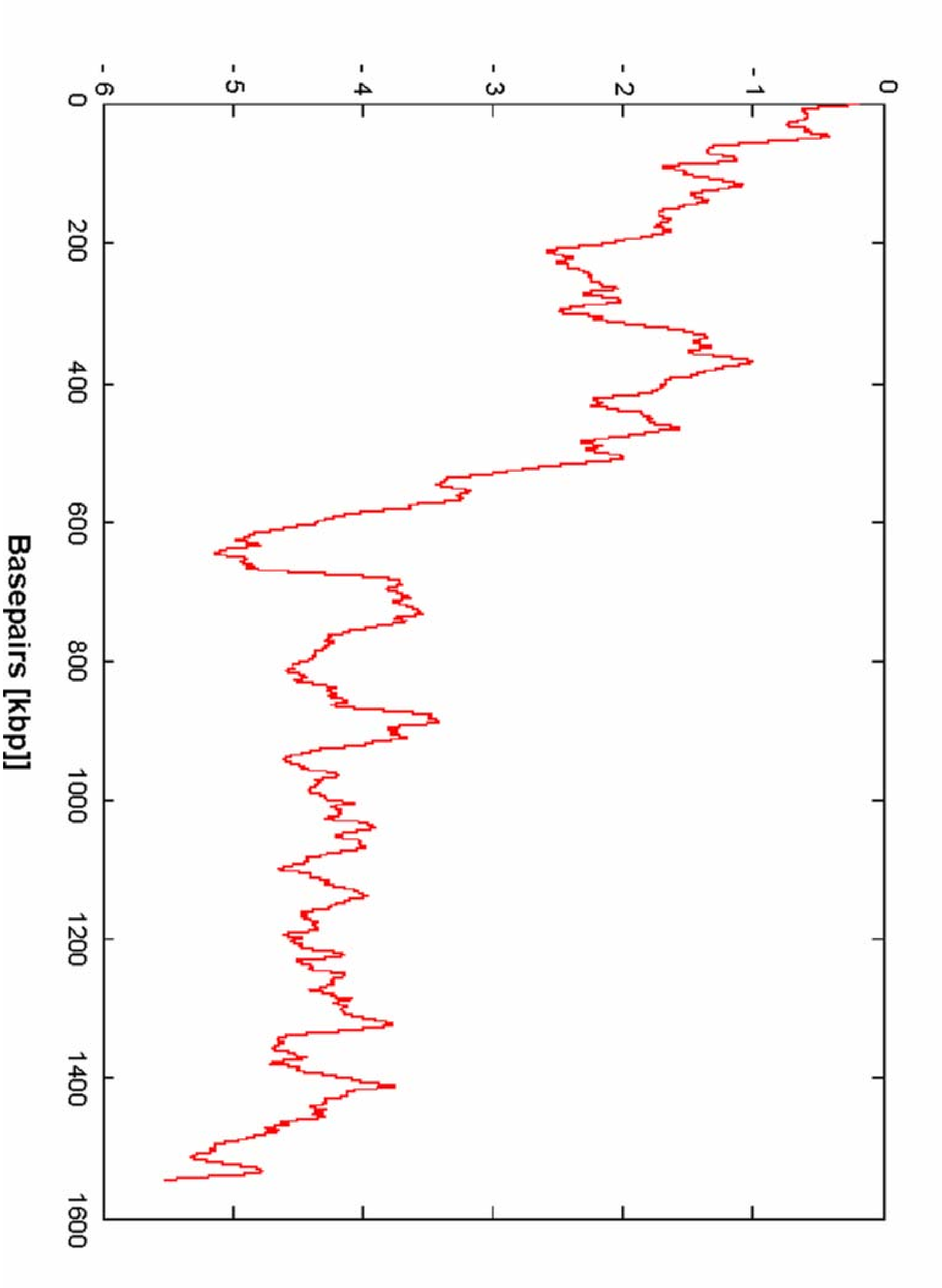
Zillig, W., Stetter, K. O., Wunderl, S., Schulz, W., Priess, H., Scholz, I. The *Sulfolobus* "*Caldariella*" group: taxonomy on the basis of the structure of DNA-dependent RNA polymerases. Arch. Mikrobiol. 1980, **125:** 259-269.

Appendix A: Keto excess and GC-skew plots of the *P. torridus* genome sequence

Keto excess plot of the *P. torridus* genome. Window = 5000bp, offset 2500 bp



Cumulative GC skew plot of the *P. torridus* genome sequence. Window = 5000bp, offset 2500 bp.

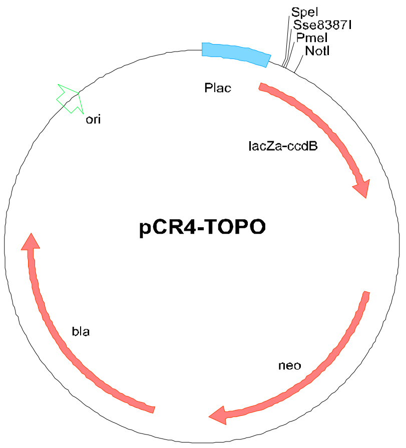


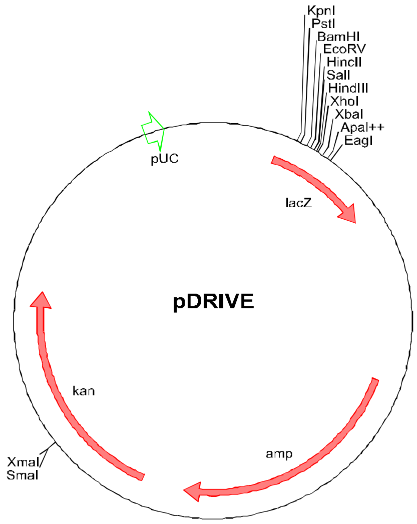
Appendix B: Codon usage table of *P. torridus*

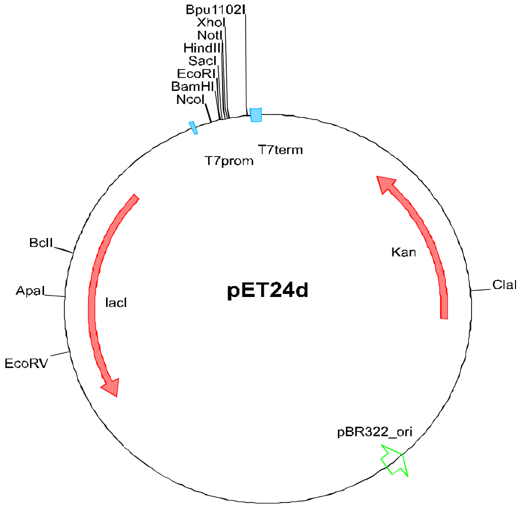
Name	Codon	Abundance	#/1000	Fraction
Gly	GGG	2421	5,98	0,08
Gly	GGA	8460	10,93	0,28
Gly	GGT	8793	23,91	0,29
Gly	GGC	10877	9,7	0,36
Glu	GAG	13802	19,11	0,52
Glu	GAA	12723	45,88	0,48
Asp	GAT	21625	37,83	0,79
Asp	GAC	5841	20,34	0,21
Val	GTG	3048	10,66	0,11
Val	GTA	7162	11,76	0,27
Val	GTT	13427	22	0,5
Val	GTC	3298	11,59	0,12
Ala	GCG	2433	6,13	0,09
Ala	GCA	14629	16,18	0,56
Ala	GCT	2611	21,09	0,1
Ala	GCC	6501	12,58	0,25
Arg	AGG	10740	9,26	0,52
Arg	AGA	7656	21,3	0,37
Ser	AGT	2479	14,17	0,07
Ser	AGC	5907	9,66	0,17
Lys	AAG	14186	30,78	0,44
Lys	AAA	18283	42,13	0,56
Asn	AAT	19627	36,02	0,63
Asn	AAC	11515	24,94	0,37
Met	ATG	14171	20,87	1
Ile	ATA	35829	17,81	0,69
Ile	ATT	10875	30,19	0,21
Ile	ATC	5106	17,09	0,1
Thr	ACG	2981	7,96	0,15
Thr	ACA	11965	17,75	0,58
Thr	ACT	1999	20,22	0,1
Thr	ACC	3610	12,58	0,18
Trp	TGG	3487	10,33	1
End	TGA	508	0,61	0,31
Cys	TGT	829	7,97	0,27

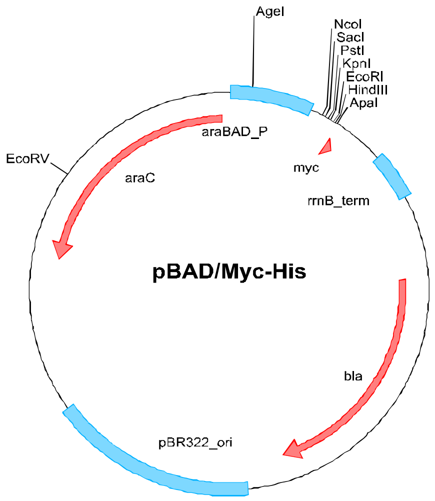
Cys	TGC	2198	4,69	0,73
End	TAG	185	0,47	0,11
End	TAA	965	0,98	0,58
Tyr	TAT	14369	18,76	0,56
Tyr	TAC	11246	14,67	0,44
Leu	TTG	3112	27,12	0,08
Leu	TTA	12064	26,37	0,3
Phe	TTT	15289	26,04	0,67
Phe	TTC	7489	18,19	0,33
Ser	TCG	2598	8,56	0,07
Ser	TCA	16296	18,79	0,47
Ser	TCT	3749	23,59	0,11
Ser	TCC	3856	14,22	0,11
Arg	CGG	264	1,74	0,01
Arg	CGA	148	3,02	0,01
Arg	CGT	1174	6,49	0,06
Arg	CGC	851	2,58	0,04
Gln	CAG	5717	12,18	0,82
Gln	CAA	1238	27,49	0,18
His	CAT	5132	13,74	0,73
His	CAC	1911	7,8	0,27
Leu	CTG	4939	10,41	0,12
Leu	CTA	3681	13,41	0,09
Leu	CTT	14463	12,17	0,36
Leu	CTC	2442	5,37	0,06
Pro	CCG	3501	5,29	0,21
Pro	CCA	7616	18,2	0,46
Pro	CCT	4002	13,59	0,24
Pro	CCC	1261	6,8	0,08

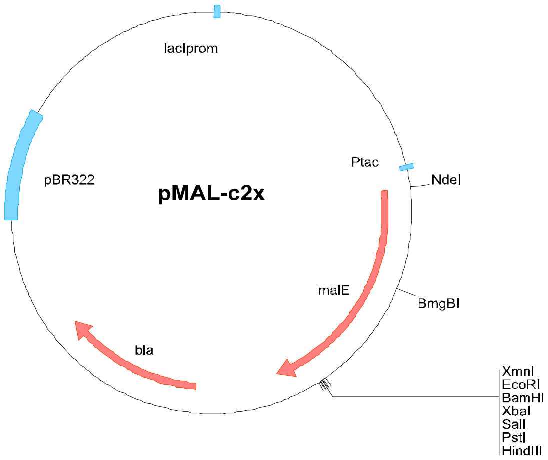
Appendix C: Plasmid maps

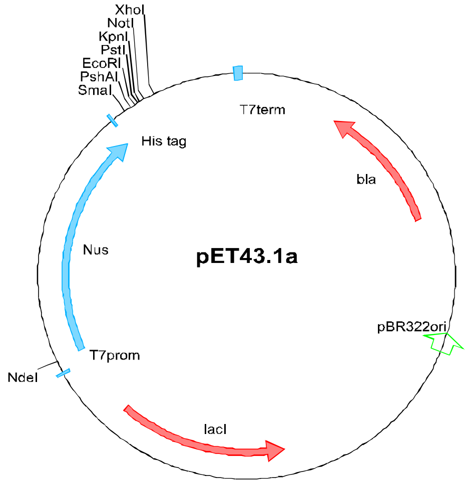


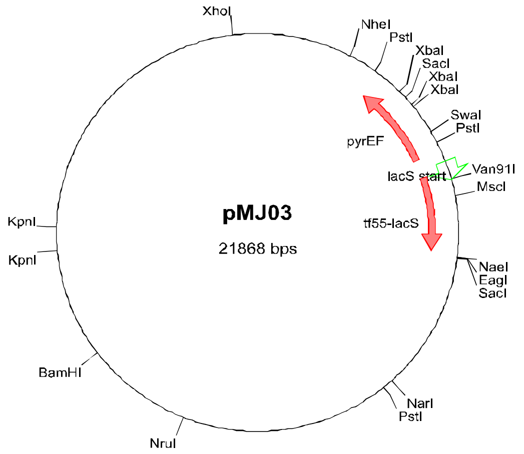


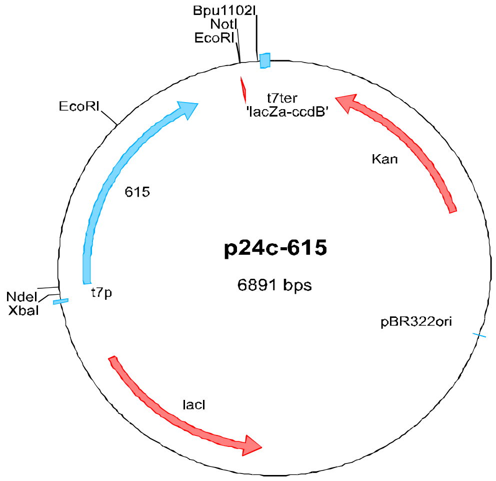


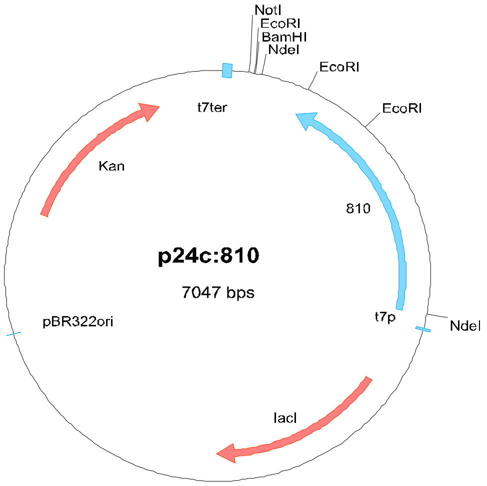

AgeI
NcoI
SacI
PstI
KpnI
EcoRI
HindIII
ApaI
EcoRV
araBAD_P
araC
myc
rrnB_term
pBAD/Myc-His
bla
pBR322_ori

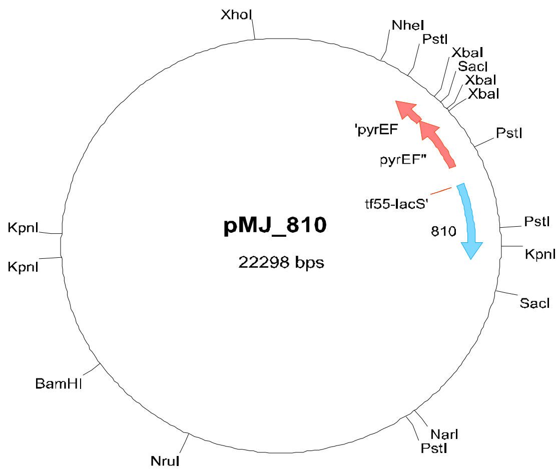

laclprom
pBR322
pMAL-c2x
Ptac
NdeI
malE
BmgBI
bla
XmnI
EcoRI
BamHI
XbaI
SalI
PstI
HindIII

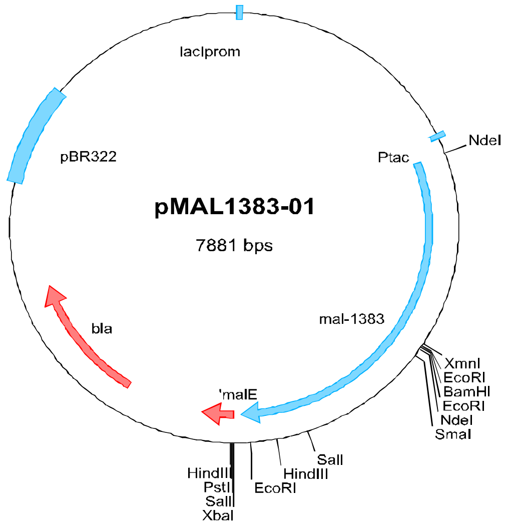

XhoI
NotI
KpnI
PstI
EcoRI
PshAI
SmaI
T7term
His tag
bla
Nus
pET43.1a
pBR322ori
T7prom
NdeI
lacI

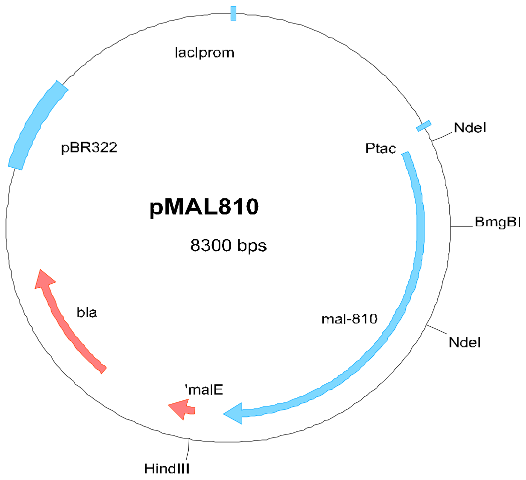

XhoI
NheI
PstI
XbaI
SacI
XbaI
XbaI
SwaI
PstI
pyrEF
Van91I
lacS start
MscI
pMJ03
21868 bps
tf55-lacS
KpnI
KpnI
NaeI
EagI
SacI
BamHI
NarI
PstI
NruI

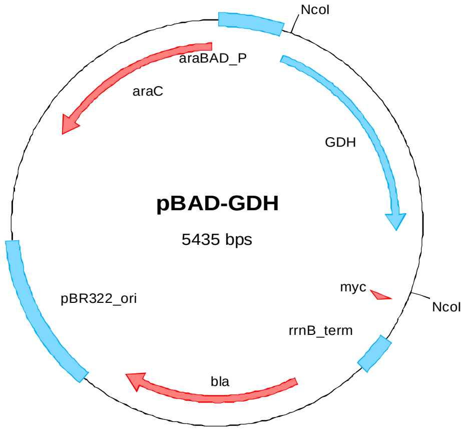

Bpu1102I
NotI
EcoRI
EcoRI
t7ter
'lacZa-ccdB'
Kan
615
p24c-615
6891 bps
NdeI
XbaI
t7p
pBR322ori
lacI

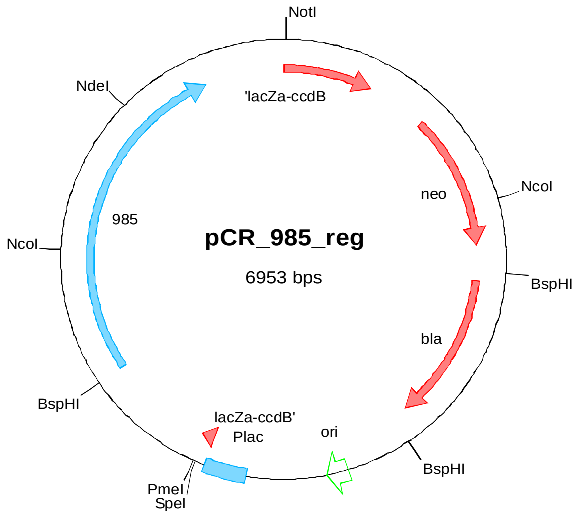

NotI
EcoRI
BamHI
NdeI
EcoRI
EcoRI
t7ter
Kan
810
p24c:810
7047 bps
pBR322ori
t7p
NdeI
lacI

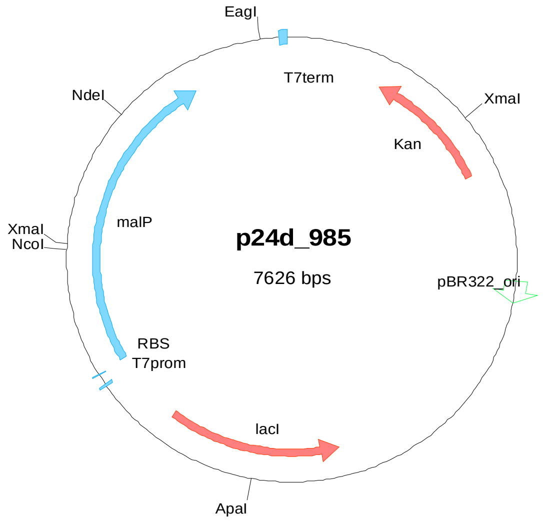

pMJ_810
22298 bps
XhoI
NheI
PstI
XbaI
SacI
XbaI
XbaI
PstI
'pyrEF
pyrEF"
tf55-lacS'
810
PstI
KpnI
SacI
NarI
PstI
NruI
BamHI
KpnI
KpnI


lacIprom
pBR322
Ptac
NdeI
pMAL1383-01
7881 bps
bla
mal-1383
XmnI
EcoRI
BamHI
EcoRI
NdeI
SmaI
'malE
SalI
HindIII
EcoRI
HindIII
PstI
SalI
XbaI


lacIprom
pBR322
Ptac
NdeI
pMAL810
8300 bps
BmgBI
bla
mal-810
NdeI
'malE
HindIII


NcoI
araBAD_P
araC
GDH
pBAD-GDH
5435 bps
myc
NcoI
pBR322_ori
rrnB_term
bla


NotI
NdeI
'lacZa-ccdB
neo
NcoI
985
NcoI
pCR_985_reg
6953 bps
BspHI
bla
BspHI
lacZa-ccdB'
Plac
ori
BspHI
PmeI
SpeI


EagI
T7term
NdeI
XmaI
Kan
XmaI
NcoI
malP
p24d_985
7626 bps
pBR322_ori
RBS
T7prom
lacI
ApaI

Appendix D: *P. torridus* ORFs with predicted signal peptides

The signal peptide prediction was done at http://www.cbs.dtu.dk/services/SignalP/. SP [%] expresses the probability of a signal peptide calculated as ($SP_{(G+)}$ + $SP_{(G-)}$ + $SP_{(Euk)}$)/3. Only probabilities above 55 % were considered significant. TMH – number of predicted trans-membrane helices.

ORF	Annotation	TMH	SP [%]
RPTO000016	membrane spanning hypothetical protein	2	58,87
RPTO000036	hypothetical SBP_bac_5	8	82,40
RPTO000039	hypothetical protein	3	99,83
RPTO000053	hypothetical protein of unknown function DUF 131	3	67,60
RPTO000071	hypothetical cytochrome oxidase assembly factor	9	58,13
RPTO000084	resistance protein	13	91,93
RPTO000097	hypothetical membrane protein	7	65,63
RPTO000124	hypothetical thermopsin precursor	4	83,13
RPTO000138	hypothetical membrane protein	3	84,87
RPTO000149	hypothetical protein	6	99,70
RPTO000157	Membrane metalloprotease	8	55,47
RPTO000166	Permease, multidrug resistance protein	13	75,63
RPTO000175	Oxidoreductase	14	55,40
RPTO000183	hypothetical membrane protein	7	67,70
RPTO000237	hypothetical membrane protein	7	63,43
RPTO000273	hypothetical protein	1	98,83
RPTO000274	hypothetical protein	2	91,47
RPTO000284	amino acid permease	15	41,13
RPTO000291	hypothetical membrane protein	3	62,30
RPTO000297	amino acid transporter	11	59,70
RPTO000326	ABC-transporter	6	60,43
RPTO000345	Malate dehydrogenase (EC 1.1.1.37)	2	96,27
RPTO000346	Succinate dehydrogenase flavoprotein subunit	3	71,60
RPTO000365	hypothetical membrane protein	5	54,73
RPTO000369	hypothetical membrane protein	5	98,03
RPTO000379	NADH-quinone oxidoreductase chain M (EC 1.6.5.3)	13	91,97
RPTO000386	ABC-transporter	7	75,83
RPTO000387	hypothetical extracellular solute binding protein	2	86,03
RPTO000401	subtilase family protein	9	96,33
RPTO000441	hypothetical	2	98,80
RPTO000455	Hypothetical Membrane Associated Protein	5	68,97
RPTO000468	hypothetical	2	66,63
RPTO000477	probable sugar transporter	10	62,70
RPTO000518	Transporter	8	45,80
RPTO000534	Extracellular solute-binding protein	5	73,53
RPTO000563	hypothetical	2	84,27
RPTO000571	probable serine protease	6	70,70
RPTO000579	hypothetical	8	42,87
RPTO000602	Hypothetical Membrane Spanning Protein	8	65,80
RPTO000606	hypothetical	5	95,83
RPTO000609	Ferredoxin	15	99,27
RPTO000624	Transporter	10	50,87
RPTO000657	Transporter	12	79,97
RPTO000659	Hypothetical Exported Protein	2	68,77
RPTO000664	Hypothetical Protein	1	71,50
RPTO000666	Transporter	7	93,53
RPTO000674	Ferrichrome-Binding Periplasmic Protein	3	79,13
RPTO000675	Ferrichrome Transport System Permease	8	61,33
RPTO000681	Hypothetical Membrane Associated Protein	2	74,07
RPTO000682	Hypothetical Membrane Associated Protein	2	83,77

ORF	Annotation	TMH	SP [%]
RPTO000683	Sodium-dependent phosphate transporter	9	99,80
RPTO000699	Permease	8	63,67
RPTO000704	Hypothetical Membrane Associated Protein	3	66,70
RPTO000706	Transporter	10	82,27
RPTO000707	Hypothetical Exported Protein	3	80,57
RPTO000748	ABC transporter, permease	6	60,63
RPTO000749	ABC transporter, permease	8	65,70
RPTO000756	NADH-quinone oxidoreductase chain A (EC 1.6.5.3)	3	61,10
RPTO000762	NADH-quinone oxidoreductase chain J (EC 1.6.5.3)	3	65,90
RPTO000763	NADH-quinone oxidoreductase chain J (EC 1.6.5.3)	2	68,13
RPTO000764	NADH-quinone oxidoreductase chain K (EC 1.6.5.3)	2	78,40
RPTO000785	Transporter, MMPL family	12	49,53
RPTO000804	Oligopeptide-binding protein	4	71,67
RPTO000806	ABC transporter, permease	7	55,70
RPTO000815	Membrane associated protein	4	95,10
RPTO000835	NADH dehydrogenase (EC 1.6.99.3)	1	62,97
RPTO000840	Amino acid permease	10	63,27
RPTO000844	THERMOPSIN PRECURSOR (EC 3.4.99.43)	4	82,97
RPTO000851	ABC transporter fusion protein	13	61,50
RPTO000852	Solute binding protein	2	40,80
RPTO000854	Hypothetical Membrane Associated Protein	3	57,43
RPTO000866	Amino acid permease	12	42,37
RPTO000867	Hypothetical Membrane Spanning Protein	3	57,13
RPTO000868	Penicillin acylase (EC 3.5.1.11)	4	64,40
RPTO000886	Transporter, MMPL family	14	86,57
RPTO000910	Extracellular solute-binding protein	5	99,23
RPTO000954	Hypothetical Membrane Associated Protein	3	94,07
RPTO000965	Transporter	9	89,77
RPTO000987	Hypothetical Membrane Spanning Protein	6	48,00
RPTO000993	Transporter	10	90,80
RPTO000995	Hypothetical Extracellular Protein	4	83,60
RPTO000998	Hypothetical Extracellular Protein	1	93,23
RPTO001031	Transporter	11	79,87
RPTO001032	Amino acid permease	11	66,90
RPTO001037	Hypothetical Membrane Associated Protein	3	54,07
RPTO001038	Potassium channel protein	3	65,70
RPTO001048	Potassium-transporting ATPase C chain (EC 3.6.3.12)	1	97,37
RPTO001054	THERMOPSIN PRECURSOR (EC 3.4.99.43)	5	99,53
RPTO001055	Permease	14	66,90
RPTO001073	Extracellular solute-binding protein	6	80,30
RPTO001074	Extracellular solute-binding protein	5	72,73
RPTO001081	Hypothetical Membrane Associated Protein	3	100,00
RPTO001102	Membrane associated Serine Protease	6	99,73
RPTO001106	Hypothetical Membrane Associated Protein	3	67,00
RPTO001127	Hypothetical Exported Protein	1	69,57
RPTO001178	Hypothetical Membrane Spanning Protein	6	57,37
RPTO001185	Hypothetical Membrane Associated Protein	9	95,50
RPTO001186	Hypothetical Exported Protein	1	99,27
RPTO001203	hypothetical membrane protein	3	91,73
RPTO001227	hypothetical membrane protein inv. in protein amidation	5	92,87
RPTO001233	Amino acid permease	11	54,30
RPTO001259	Glucoamylase (EC 3.2.1.3)	6	92,77
RPTO001261	hypothetical membrane protein	7	88,77
RPTO001275	hypothetical protein inv. in cytochrome c biogenesis	5	99,27
RPTO001345	dehydrogenase	4	68,37
RPTO001346	Permease, multidrug resistance protein	14	98,33
RPTO001356	hypothetical sugar transporter	9	64,37
RPTO001372	hypothetical protein	2	99,90

ORF	Annotation	TMH	SP [%]
RPTO001386	hypothetical protein	2	68,10
RPTO001393	hypothetical ATP synthase subunit C	2	98,03
RPTO001397	hypothetical protein	2	60,67
RPTO001399	hypothetical protein	1	68,93
RPTO001412	Mechanosensitive (MS) ion channel	4	64,13
RPTO001414	Hypothetical Membrane Associated Protein	3	78,57
RPTO001420	Molybdopterin biosynthesis MoeB protein	2	43,13
RPTO001466	Hypothetical Membrane Spanning Protein	3	62,63
RPTO001470	Hypothetical Exported Protein	4	98,47
RPTO001495	Multiple antibiotic resistance protein marC	5	85,60
RPTO001500	Shikimate kinase (EC 2.7.1.71)	2	78,80
RPTO001503	Hypothetical Exported Protein	3	97,63
RPTO001508	Cytochrome d ubiquinol oxidase subunit I (EC 1.10.3.-)	9	71,23
RPTO001529	Transporter	8	58,20
RPTO001533	Transporter involved in lipid transport	13	70,70

Acknowledgements

First of all, I would like to thank Prof. Dr. Liebl for giving me the opportunity to work in his group and for his willingness to share with me his knowledge and expertise. Also, I greatly appreciate his help during the preparation of this thesis.

I am grateful to Prof. Dr. Gottschalk for his readiness to be a co-referent of this thesis.

It was a big pleasure for me to take part in the *P. torridus* genome project with Dr. Fütterer and I highly appreciate his contribution to the present work. In addition, I want to thank all the members of the project who keep it running: Prof. Gottschalk, Dr. H. Liesegang, C. Schepers, B. Dock and Jessica Jabin.

I am much indebted to Dr. Tzvetkov for supporting me during the whole time of my work, for the many constructive discussions, and for the thorough review of this manuscript that helped to substantially decrease the number of inconsistencies in the text.

I want to thank all my current and former colleagues from the working group of Prof. Liebl / Prof. Streit: Martin Armbrecht, Meike Ballschmiter, Jörg Kleine, Volker Hoffmann, Christel Schmeißer, Krasimira Ivanova, as well as Christian Elend, Sonja Voget, Ute Ludwig, Elke Heinz etc. for creating a pleasant atmosphere in the lab and helping me throughout my work.

Special tanks to Meike Ballschmiter and Jorg Kleine for the plenty of vital discussions and for the critical reading of the chapters.

I ought to mention here the hospitality of Prof. Schleper and Melanie Jonuscheit and their willingness in showing me how transformation of *S. solfataricus* works.

The bulgarian community in Göttingen has been very supportive during my work on this project and I am very much obliged.

Finally, I want to thank my wife for her patience, understanding and support during the time spent in the preparation of this work.

Lebenslauf

Persönliche Daten

Name	Angel Angelov
27.06.1974	geboren in Botevgrad, Bulgarien Staatsangehörigkeit: bulgarisch

Schulausbildung, wissenschaftlicher Werdegang

1981 - 1988	Besuch der GrundschuleBotevgrad, Bulgarien
1988 - 1993	Besuch des Gymnasium „Aleko Konstantinov“ Pravez, Bulgarien
Juni 1993	Allgemeine Hochschulreife
Oktober 1993	Immatrikulirung an der „St. Kliment Ohridski“ Universität, Sofia für den Studiengang Molekularbiologie (Magister)
1997-1999	Spezialisierung im Fach Genetik in der Abteilung Prenataldiagnostik an der Medizinische Hochschule Sofia
Juli1998 - Juli 1999	Anfertigung der experimentellen Diplomarbeit unter Anleitung von Dr. Savov mit dem Titel: „Untersuchung der Häufigkeit der genetisch bedingten Resistenz gegen HIV-1-Infektion in der bulgarischen Population“
Januar 2000-September 2000	Militärdienst, Vratza, Bulgarien
Januar 2001 - September 2001	Tätig als wiss. Mitarbeiter im Nationalen-HIV-Labor, Sofia
November 2001	Beginn der experimentellen Arbeiten zur vorliegenden Dissertation