Computational Genomics of Hyperthermophiles

Harmen J.G. van de Werken

Promotoren

Prof. dr. J. van der Oost
Persoonlijk hoogleraar bij de leerstoelgroep Microbiologie
Wageningen Universiteit

Prof. dr. W.M. de Vos
Hoogleraar in de Microbiologie
Wageningen Universiteit

Copromotor

Dr. S.W.M. Kengen
Universitair docent bij de leerstoelgroep Microbiologie
Wageningen Universiteit

Promotiecommissie

Prof. dr. J.A.M. Leunissen
Wageningen Universiteit

Prof. dr. O.P. Kuipers
Rijksuniversiteit Groningen

Prof. dr. R.J. Siezen
Radboud Universiteit Nijmegen

Dr. B. Snel
Universiteit Utrecht

Dit onderzoek is uitgevoerd binnen de onderzoekschool VLAG

Computational Genomics of Hyperthermophiles

Harmen J.G. van de Werken

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
Prof. dr. M.J. Kropff,
in het openbaar te verdedigen
op woensdag 16 april 2008
des namiddags te vier uur in de Aula.

Computational Genomics of Hyperthermophiles
Harmen J.G. van de Werken

Ph.D. thesis Wageningen University, Wageningen, the Netherlands (2008)
151 pages, with summary in Dutch

ISBN 978-90-8504-894-7

Abstract

With the ever increasing number of completely sequenced prokaryotic genomes and the subsequent use of functional genomics tools, *e.g.* DNA microarray and proteomics, computational data analysis and the integration of microbial and molecular data is inevitable. This thesis describes the computational analyses on (hyper)thermophilic archaeal and bacterial genomes with a particular emphasis on carbohydrate metabolic pathways and their regulation. These analyses were integrated with wet-lab functional genomics data and results from classical molecular biology and microbial physiology experiments. The research was conducted on the archaea *Sulfolobus solfataricus*, *Pyrococcus furiosus*, *Thermococcus kodakaraensis* and the hydrogen producing bacterium *Caldicellulosiruptor saccharolyticus.*

The reconstruction of the central carbohydrate metabolism in the thermo-acidophile *S. solfataricus* was carried out by a combination of genome sequence, whole transcriptome and proteome analyses. Only slight differences in the mRNA and the protein expression levels were detected when *S. solfataricus* was grown on peptides vs. glucose. However, the breakdown of D-arabinose vs. D-glucose revealed a complete novel pathway in the domain of Archaea. Similar catabolic pathways were identified in other prokaryotes and therefore a comprehensive genomic reconstruction was carried out on the pentose utilizing pathways in Archaea and, additionally, the results were compared to Bacteria and Eukarya.

A computational promoter analysis of the glycolytic genes in the anaerobic species of the order Thermococcales (*P. furiosus* and *T. kodakaraensis*) indicated a clear *cis*-regulatory element that putatively controls all the genes of the glucose and starch degrading pathways. A comparative genomic analysis of the hyperthermophilic Thermococcales species led to the discovery of a putative transcriptional regulator that is probably involved in regulation of the entire regulon.

The complete genome sequence of the extremely thermophilic *Caldicellulosiruptor saccharolyticus* revealed a circular genome of 2,970,275 base pairs that encodes 2679 putative proteins. The central carbohydrate pathways of *C. saccharolyticus* were studied in detail and the pathways for producing biohydrogen from plant cell wall material were unraveled. Subsequently, a whole transcriptome analysis of *C. saccharolyticus* grown on different monosaccharides showed a tight transcriptional regulation of these pathways, without glucose-based catabolite repression. *C. saccharolyticus* is therefore a good candidate to produce molecular hydrogen from biomass feedstock.

The new insights into how prokaryotic genomes, genes and their encoded proteins function, as described in this thesis, can be applied on hyperthermophilic proteins and strains for use in and improvement of industrial processes.

Table of contents

Chapter 1

General Introduction

Hyperthermophiles

Hyperthermophiles are organisms that grow optimally at temperatures above 80 ºC, whereas thermophiles have an optimum growth temperature between 60 ºC and 80 ºC (Brock and Freeze, 1969) (Stetter, 1996). Organisms that thrive at these high temperatures are all prokaryotes (Archaea and Bacteria), while no (hyper)thermophilic eukaryotes (Eukarya) have been found to date. Until the late 1960s it was generally assumed that 55 ºC was the upper temperature of life. However, in thermal springs of Yellowstone National Park, Thomas Brock discovered *Thermus aquaticus*, an organism that grows optimally at 70 ºC and with a maximum of 79 ºC (Brock and Freeze, 1969). Subsequently, many more organisms were discovered with even higher optimal and maximal growth temperatures, *e.g. Sulfolobus acidocaldarius* (Brock *et al.*, 1972). Initially, thermophilic organisms were mainly isolated from terrestrial ponds, but subsequently species were isolated from marine environments. After the discovery of deep-sea vents in the 1970s, the most extreme hyperthermophiles described to date were found in these hydrothermal systems. Examples are *Pyrolobus fumarii*, isolated from a black smoker wall at the Mid Atlantic Ridge, is able to grow at 113 ºC with an optimal temperature of 106 ºC (Blochl *et al.*, 1997) and, an iron-reducing strain that was isolated from the Pacific Ocean which has been reported to grow even at 121 ºC (Kashefi and Lovley, 2003).

Most of the (hyper)thermophiles belong to the domain of the Archaea (Fig. 1.1). The classification of Archaea was introduced by Carl Woese and was based on comparative analyses of ribosomal RNA (rRNA) sequences (Woese and Fox, 1977; Woese *et al.*, 1990). He discovered that prokaryotes, unicellular organisms that lack a nucleus, can be divided into two distinct evolutionary groups: the Bacteria and the Archaea. Together with the Eukarya domain, these three domains cover all the life forms on earth.

Besides studying these fascinating extreme life forms at the upper temperature limit of life, hyperthermophilicity has been studied intensively in order to reveal the molecular basis of thermostability of biological macromolecules (DNA, RNA, proteins). At the level of DNA, the gene coding for reverse gyrase has been regarded as a molecular marker of hyperthermophilicity; the reverse gyrase enzyme induces positive supercoiling of DNA (Kikuchi and Asai, 1984). However, reverse gyrase has also been detected in thermophilic bacteria such as *Thermus thermophilus*, *Thermoanaerobacter tengcongensis* and *Caldicellulosiruptor saccharolyticus* (Brochier-Armanet and Forterre, 2007). These microbes have an optimal growth temperature (T_{opt}) between 70 and 75 ºC (Table 1.1) and can, therefore, be regarded as borderline-hyperthermophiles. Moreover, the recently sequenced thermophilic ε-proteobacteria *Caminibacter mediatlanticus* (T_{opt} 55 ºC (Voordeckers *et al.*, 2005)) and *Nitratiruptor* sp. (T_{opt} 55 ºC (Nakagawa *et al.*, 2007)) also possess reverse gyrase. Thus, reverse gyrase is not a universal marker for hyperthermophilicity, although all hyperthermophiles possess the reverse gyrase gene.

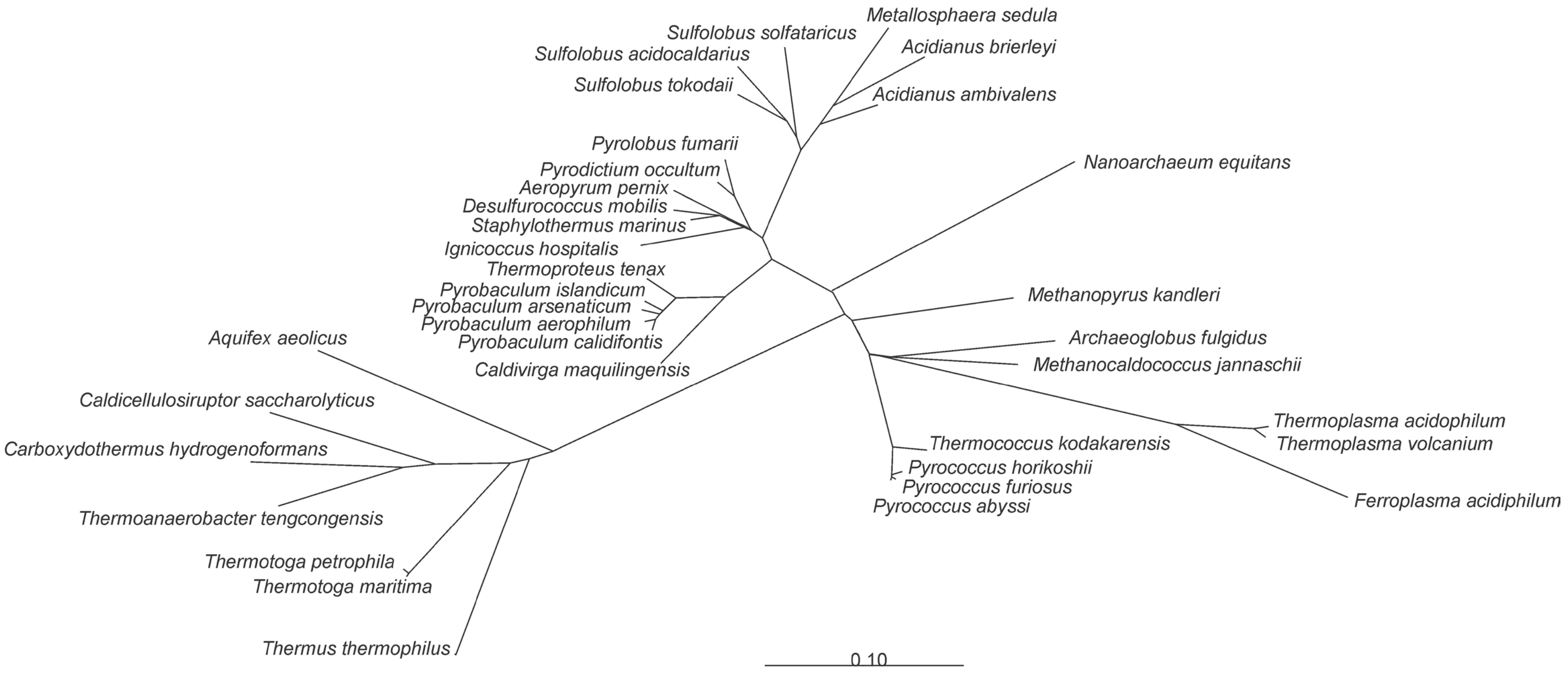


Figure 1.1 Phylogenetic tree based on (hyper)thermophilic archaeal and bacterial SSU rRNA sequences. Alignment and phylogenetic analysis were performed with the ARB software (Ludwig *et al.*, 2004), and the tree was constructed using the neighbor-joining method (Saitou and Nei, 1987). The reference bar indicates the branch length that represents 10% diversity.

Transfer RNAs, ribosomal RNAs and other non-coding RNAs (ncRNAs) have a GC-content that correlates with the growth temperature of the organism (Galtier and Lobry, 1997). Since the GC pair is more stable than the AT pair due to an extra hydrogen bond, hyperthermophiles have more stable ncRNAs. In contrast, no correlation between the growth optimum and the GC-content of the completely sequenced genomes has been detected and therefore messenger RNAs from hyperthermophiles do not contain a higher GC-content.

Proteins of hyperthermophiles are more stable at elevated temperatures and are more resistant to chemical denaturants compared to mesophilic counterparts; however, hyperthermophilic proteins are often not active under mesophilic conditions. Several studies comparing hyperthermophilic proteins to the mesophilic structural homologs, reveal that different combinations of distinct stabilizing strategies contribute to enhanced protein stability (reviewed by (Daniel *et al.*, 2008; Eijsink *et al.*, 2005)). A general feature appears to be the fact that hyperthermophiles contain a larger number of charged residues at the surface of the proteins, potentially stabilizing the protein through ion-pair formation (Cambillau and Claverie, 2000).

Hyperthermophilic enzymes with commercial applications are used in molecular biology, starch processing and other biotechnological and industrial processes (Vieille and Zeikus, 2001). Ever since the heat-resistant DNA-polymerase of *T. aquaticus* (Taq polymerase) was used for the polymerase chain reaction (PCR) (Saiki *et al.*, 1988), both the scientific world and the industry have great interest in hyperthermophilic enzymes (reviewed by (Atomi, 2005; Unsworth *et al.*, 2007)). Whole-cell applications of hyperthermophiles are, however, uncommon. Nevertheless, considerable progress has been made in developing genetic tools that could be used for metabolic engineering of these heat-loving microbes (Sato *et al.*, 2003).

Metabolism of hyperthermophiles

Hyperthermophiles display a high metabolic diversity. They are able to grow fermentatively (*Pyrococcus*, *Thermotoga*), but are also able to respire aerobically (*Sulfolobus*) and anaerobically (*Pyrobaculum*) and obtain carbon from organic molecules (heterotrophs: *Thermococcus*) or from CO_2 (autotrophs: *Methanocaldococcus*) (Schonheit and Schafer, 1995). However, a hyperthermophilic phototroph has never been found. The carbohydrate metabolism of hyperthermophiles, in particular the catabolism of monosaccharides has been studied intensively. The Embden-Meyerhof (EM) and the Entner–Doudoroff (ED) catabolic pathways are similar to the mesophilic counterparts, but have unique conversions, novel enzymes and distinct regulatory mechanisms (for reviews of archaeal carbohydrate metabolism: (Siebers and Schonheit, 2005; Verhees *et al.*, 2003)). Additionally, enzymes that are able to hydrolyze glycosidic bonds in polysaccharides (glycoside hydrolases) are widely distributed in hyperthermophiles.

Table 1.1 Completely sequenced genomes of hyperthermophiles and borderline-hyperthermophiles in the Genomes Online Database (GOLD) (Liolios *et al.*, 2007).

Species	Strain	Lifestyle		T_{opt} (°C)^a	Genome size (kbp)	Proteins	GC-content (%)	Chromosomes	Plasmids	Reference
Archaea										
Aeropyrum pernix	K1	AE	H	90	1669	1700	56.3	1	0	(Kawarabayasi *et al.*, 1999)
Archaeoglobus fulgidus	VC-16	AN	FA	83	2178	2420	48.6	1	0	(Klenk *et al.*, 1997)
Caldivirga maquilingensis	IC-167	AE	H	85	2077	1963	43	1	0	Unpublished
Hyperthermus butylicus	DSM 5456	AN	H	101	1667	1602	53	1	0	(Brugger *et al.*, 2007)
Ignicoccus hospitalis	Kin4/I	AN	A	90	1297	1434	61.2	1	0	Unpublished
Metallosphaera sedula	DSM 5348	AE	FA	75	2191	2256	46.3	1	0	(Auernik *et al.*, 2007)
Methanocaldococcus jannaschii	DSM 2661	AN	A	85	1664	1729	31.4	1	2	(Bult *et al.*, 1996)
Methanopyrus kandleri	AV19	AN	A	98	1694	1687	61.2	1	0	(Slesarev *et al.*, 2002)
Nanoarchaeum equitans	Kin4-M	AN	P	90	490	536	31.6	1	0	(Waters *et al.*, 2003)
Pyrobaculum aerophilum	IM2	FAN	FA	100	2222	2605	51.4	1	0	(Fitz-Gibbon *et al.*, 2002)
Pyrobaculum arsenaticum	PZ6	AN	FA	68–100	2121	2298	58.3	1	0	Unpublished
Pyrobaculum calidifontis	JCM 11548	FAN	H	90 - 95	2009	2149	57.2	1	0	Unpublished
Pyrobaculum islandicum	DSM 4184	AN	FA	100	1826	1978	49.6	1	0	Unpublished
Pyrococcus abyssi	GE5	AN	H	96	1765	1896	44.7	1	1	(Cohen *et al.*, 2003)
Pyrococcus furiosus	JCM 8422	AN	H	100	1908	2125	40.8	1	0	(Robb *et al.*, 2001)
Pyrococcus horikoshii	OT3	AN	H	98	1738	1955	41.9	1	0	(Kawarabayasi *et al.*, 1998)
Pyrolobus fumarii		FAN	A	106	1850	2000	53			proprietary genome sequence
Staphylothermus marinus	F1	AN	H	92	1570	1570	35	1	0	Unpublished
Sulfolobus acidocaldarius	DSM 639	AE	H	80	2225	2292	36.7	1	0	(Chen *et al.*, 2005)
Sulfolobus solfataricus	P2	AE	H	80	2992	2977	35.8	1	0	(She *et al.*, 2001)
Sulfolobus tokodaii	7, JCM 10545	AE	H	80	2694	2825	32.8	1	0	(Kawarabayasi *et al.*, 2001)
Thermococcus kodakaraensis	KOD1	AN	H	85	2088	2306	52	1	0	(Fukui *et al.*, 2005)
Thermofilum pendens	Hrk 5	AN	H	88	1781	1824	57.7	1	1	Unpublished
Bacteria										
Aquifex aeolicus	VF5	FAN	A	95	1551	1529	43.5	1	1	(Deckert *et al.*, 1998)
Caldicellulosiruptor saccharolyticus	DSM 8903	AN	H	70	2970	2679	35	1	0	This thesis Chapter 5
Carboxydothermus hydrogenoformans	Z-2901	AN	A	78	2401	2620	42	1	0	(Wu *et al.*, 2005)
Thermoanaerobacter tengcongensis	MB4T / JCM 11007	AN	H	75	2689	2588	37.6	1	0	(Xue *et al.*, 2001)
Thermotoga maritima	MSB8	AN	H	80	1860	1858	46.2	1	0	(Nelson *et al.*, 1999)
Thermotoga petrophila	RKU-1	AN	H	80	1823	1785	46	1	0	Unpublished
Thermus thermophilus	HB8	AE	H	75	1849	1973	69.4	1	2	Unpublished
Thermus thermophilus	HB27	AE	H	75	1894	1982	66.6	1	1	(Henne *et al.*, 2004)

AE, aerobe; AN, anaerobe; FAN, facultative anaerobe; H, heterotroph; A, autotroph; FA, facultative autotroph; P, parasite.

[a] $T_{opt:}$ optimal growth temperature or temperature growth range of microbes according to GOLD database, Prokaryotic Growth Temperature database (PGTdb) (Huang *et al.*, 2004) or species description.

Extracellular and intracellular hyperthermozymes are capable of hydrolyzing the α-glycosidic bond or the β-glycosidic bond of the polymers which can be further metabolized to the final end products, such as acetate, lactate and CO_2. The significant number of glycoside hydrolases in for instance *Thermotoga maritima* and *Sulfolobus* reflects their saccharolytic capabilities

Transcriptional regulation of hyperthermophiles

Protein synthesis can be modulated by activation or repression of the transcription of DNA to RNA, or of the translation of RNA to protein. Moreover, altering protein activity by allosteric regulation or post-translational modifications and degradation of protein are ubiquitous processes in the prokaryotic cell. Transcriptional regulation is a major control point of gene expression in prokaryotes. The prokaryotic domains, Archaea and Bacteria have a different type of RNA polymerase (RNAP). The bacterial RNAP consists of four different subunits in the stoichiometry $\alpha_2\beta\beta'\omega$ and is associated with one sigma factor (σ-factor) forming the RNAP holoenzyme. The archaeal RNAP is, however, more similar to the eukaryal RNAP II than to the bacterial counterpart. It consists of 10 to 12 subunits and requires two general transcription factors for initiating transcription: transcription factor B (TFB) and TATA-binding protein (TBP) (Bell *et al.*, 2001).

Despite the resemblance of the RNAP between Archaea and Eukarya, most of the transcriptional regulators from Archaea belong to the bacterial families of regulators (Aravind and Koonin, 1999). Hence, most of the transcriptional regulators in hyperthermophiles are bacterial-types and only a few are biochemically characterized (reviewed by (Bell, 2005; Brinkman *et al.*, 2003; Geiduschek and Ouhammouch, 2005)).

Computational genomics

Since the publication of the first complete genome sequence of the free-living organism *Haemophilus influenzae* (Fleischmann *et al.*, 1995) and that of the first hyperthermophilic archaeon *Methanocaldococcus jannaschii* (Bult *et al.*, 1996), a wealth of sequence data has emerged. At the time of writing, 25 hyperthermophilic genomes were completely sequenced (Table 1.1) and many sequencing projects are ongoing (Liolios *et al.*, 2007).

To distill the actual biology from the complete genome sequence, the use of computational analysis, *i.e.* computational genomics, is inevitable (Koonin, 2001). Computational genomics not only comprises assembling genomes, predicting genes and identifying regulatory motifs at the DNA level, but also includes sequence alignments (genes and proteins), phylogenetic analysis and protein function prediction by, for example, comparative genomics.

Gene prediction in prokaryotic genomes can be very accurate using computational programs, such as Glimmer (Delcher *et al.*, 2007) and Critica (Badger and Olsen, 1999).

However, despite the insights that genome sequencing projects give and the accurate gene/ORF predictions these programs generate, function prediction is still a complicated matter. Functions are unknown for many encoded proteins and for some proteins the function is only generally known. The recent completely sequenced extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus* with 2679 predicted open reading frames, for instance, has 875 genes that code for proteins with unknown function and 225 genes with general function only. Since many of these genes will never be biochemically characterized or experimentally studied, computational methods have been developed. New tools should be developed to improve function prediction of these uncharacterized proteins and genes.

Function prediction of proteins can be carried out by two fundamentally different methods (reviewed by (Ettema *et al.*, 2005; Gabaldón and Huynen, 2004)). First, sequence similarity based function prediction and second, genomic-context based function prediction. Sequences similarity detection methods are based on the fact that proteins from a common ancestor do have a similar function and, in particular, the molecular function of a characterized protein can therefore be copied to the protein of interest. Algorithms such as BLAST (Altschul *et al.*, 1990), Smith-Waterman (Smith and Waterman, 1981) and FASTA (Pearson, 1990) were constructed to detect similarities by searching databases of protein sequences. Subsequently, more sensitive profile-based algorithms, *e.g.* PSI-BLAST (Altschul *et al.*, 1997) and HMMER (Eddy, 1998), were developed. Second, ”Genomic-context” function prediction methods refers to a more comparative genomics approach, aiming at elucidating the involvement of the protein in a certain biological process. The STRING database (von Mering *et al.*, 2007) integrates many of the genomic-context methods such as: (1) gene neighborhood conservation, (2) gene fission or fusion events and, (3) similar and complementary phylogenetic distribution of proteins. The database combines these findings with experimental co-expression data and protein-protein interaction data with the aim to improve the protein function predictions. Moreover, co-evolution of sequences and regulon predictions can generate useful information on a protein’s function.

Function and gene prediction of small ncRNAs is a complicated task, since these genes are relatively small and do not have a clear start and end codon as in protein-coding genes. Algorithms for predicting tRNAs (Lowe and Eddy, 1997) and small nucleolar RNAs (Lowe and Eddy, 1999) are very accurate to date. Furthermore, Rfam database provides aligned ncRNA families and a sequences search algorithm INFERNAL (Griffiths-Jones *et al.*, 2005). Nevertheless, *de novo* small ncRNAs prediction is a laborious task and new algorithms are necessary to improve on precision and accuracy of ncRNA gene prediction. However, in hyperthermophiles, due to the high GC-content of ncRNA, novel genes have been predicted and their transcripts have been measured (Klein *et al.*, 2002). Additionally, a defense mechanism has recently been detected in prokaryotes (Barrangou *et al.*, 2007). This prokaryotic defense system contains the so-called clustered regularly interspaced short palindromic repeats (CRISPR) and includes CRISPR-associated (CAS) proteins. The ncRNA CRISPR-transcripts are used to detect and (most likely) to degrade foreign DNA or RNA, which is mediated by the CAS proteins and gives the cell immunity against viruses. In addition, it has been postulated that the CRISPR-system could also function as an analogous eukaryotic RNA-interference system and therefore be involved in regulating gene expression. Several tools have been developed to predict the occurrence of the CRISPR-system and the array of non-coding genes (Bland *et al.*, 2007; Edgar, 2007; Grissa *et al.*, 2007). Remarkably, CRISPRs are detected in all hyperthermophiles sequenced to date.

Regulatory sequences such as DNA-binding motifs, which are even smaller than ncRNAs, have been predicted in genomes by clustering of upstream regions of functionally related genes or orthologous genes (phylogenetic footprinting). Motif predicting algorithms, such as MEME (Bailey and Elkan, 1994) and Gibbs Recursive Sampler (Thompson *et al.*, 2003) can analyze the DNA sequences. Subsequently, DNA sequences can be searched with the motif to find additional *cis*-regulatory elements. The objective of these analyses is to predict genes that are under control of the same transcriptional regulator (regulon) or global regulator (modulon) and to describe the regulatory networks, but it can also be useful as tool for function prediction.

High-throughput post-genomic technologies

Besides classical approaches to study the physiology, biochemistry and molecular biology of the (hyper)thermophiles, the application of high-throughput methods is an interesting alternative, in particular, when the complete genome of the organism is available. Depending on the target molecules, the high-throughput methods can be divided into (1) transcriptomics for RNA, (2) proteomics and structural genomics for proteins and, (3) metabolomics for metabolites. Computational methods are necessary to analyze and interpret the data from high-throughput systems, because the amount of data generated is overwhelming. The high-throughput study of functions and interactions of biomolecules is commonly described as functional genomics, whereas computational molecular biology, using data from different techniques, is called integrated genomics.

Several studies have analyzed the whole transcriptome of hyperthermophilic species. Growing the archaeon *P. furiosus* (Schut *et al.*, 2003; Weinberg *et al.*, 2005) or bacterium *Thermotoga maritima* (Nguyen *et al.*, 2004) on different carbon sources or under different stress conditions demonstrated the complexity and diversity of the transcriptional regulatory networks. In addition, proteomics studies revealed the differential expression of many genes and confirmed the complicated biology in these microbes *e.g. Methanocaldococcus jannaschii* (Zhu *et al.*, 2004). Since proteins from hyperthermophiles are easy to purify when using a mesophilic host, structural genomics projects have been undertaken for *T. maritima* (DiDonato *et al.*, 2004)*, P. furiosus* (Adams *et al.*, 2003) and *T. thermophilus* (Yokoyama *et al.*, 2000). A large number of hyperthermophilic 3-D protein structures are now deposited in structural databases. Finally, metabolomics, the study to quantify the metabolites in a cell, is an emerging field in hyperthermophiles.

Aim and outline of this thesis

This thesis comprises genome analyses that predict gene function, unravel metabolic pathways, describe gene expression, and improve insights in evolution of prokaryotes. The key objective is to understand how genomes actually function, providing a basis for future applications in industry by improving hyperthermophilic strains and proteins. A general theme in the described research considers the computational analysis of metabolism in hyperthermophiles, and the integration with high-throughput functional genomics and classical molecular biological approaches. Three different approaches have been used aiming at carbohydrate-converting pathways and their regulation: (i) genome sequence analysis, (ii) expression data analysis at mRNA and protein level, and (iii) integrated genomics (*in silico*, *in vitro*, *in vivo*).

Chapter 1 gives an overview of the characteristics of hyperthermophiles and, in particular, their carbohydrate metabolism, transcriptional regulation, and phylogeny. Besides the intriguing biology of hyperthermophiles the chapter summarizes and discusses the hyperthermophilic genomes completely sequenced to date. Moreover, it recapitulates the computational and functional tools that have been employed on hyperthermophiles and their genomes. These analyses are primarily focused on unraveling protein function, reconstructing metabolism and describing regulation at the transcriptional level.

Chapter 2 describes the reconstructed central carbohydrate metabolism of *Sulfolobus solfataricus*. Expression data of genes and proteins involved in these central pathways were determined and analyzed. To our surprise only small differences were detected when comparing *S. solfataricus* grown on either peptides or glucose.

Chapter 3 reviews the archaeal pentose metabolism and compares it to Bacteria and Eukarya. This review gives insight into the evolutionary and versatility of pentose anabolism and catabolism in Archaea.

Chapter 4 describes a comprehensive analysis of the carbohydrate metabolism in *Thermococcus kodakaraensis* and *Pyrococcus furious*. It reveals the largest archaeal regulon described to date by a combination of bioinformatics and experimental analyses. An identified *cis*-acting element and a putative regulator are predicted to be involved in the regulation at the transcriptional level of carbohydrate metabolism in both species.

Chapter 5 describes the complete genome sequence of the hydrogen producer and extremely thermophilic organism *Caldicellulosiruptor saccharolyticus*. The central saccharolytic pathways for producing biohydrogen from plant cell wall material are unraveled. Moreover, transcriptome analyses reveal the response of *C. saccharolyticus* growing on different carbohydrates.

Chapter 6 summarizes this thesis, confers the obtained results and concludes with future perspectives. Additionally, it compares the valuable computational predictions to recent experimental data and discusses the advantages and limitations of computational genomics.

References

Adams, M. W., Dailey, H. A., DeLucas, L. J., Luo, M., Prestegard, J. H., Rose, J. P., and Wang, B. C. (2003). The Southeast Collaboratory for Structural Genomics: a high-throughput gene to structure factory. Acc Chem Res *36*, 191-198.

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

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Aravind, L., and Koonin, E. V. (1999). DNA-binding proteins and evolution of transcription regulation in the archaea. Nucleic Acids Res *27*, 4658-4670.

Atomi, H. (2005). Recent progress towards the application of hyperthermophiles and their enzymes. Curr Opin Chem Biol *9*, 166-173.

Auernik, K. S., Maezato, Y., Blum, P. H., and Kelly, R. M. (2007). Genome sequence of the metal-mobilizing, extremely thermoacidophilic archaeon *Metallosphaera sedula* provides insights into bioleaching metabolism. Appl Environ Microbiol.

Badger, J. H., and Olsen, G. J. (1999). CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol *16*, 512-524.

Bailey, T. L., and Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc Int Conf Intell Syst Mol Biol *2*, 28-36.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709-1712.

Bell, S. D. (2005). Archaeal transcriptional regulation--variation on a bacterial theme? Trends Microbiol *13*, 262-265.

Bell, S. D., Magill, C. P., and Jackson, S. P. (2001). Basal and regulated transcription in Archaea. Biochem Soc Trans *29*, 392-395.

Bland, C., Ramsey, T. L., Sabree, F., Lowe, M., Brown, K., Kyrpides, N. C., and Hugenholtz, P. (2007). CRISPR recognition tool (CRT): a tool for automatic detection of clustered regularly interspaced palindromic repeats. BMC Bioinformatics *8*, 209.

Blochl, E., Rachel, R., Burggraf, S., Hafenbradl, D., Jannasch, H. W., and Stetter, K. O. (1997). *Pyrolobus fumarii*, gen. and sp. nov., represents a novel group of archaea, extending the upper temperature limit for life to 113 degrees C. Extremophiles *1*, 14-21.

Brinkman, A. B., Ettema, T. J., de Vos, W. M., and van der Oost, J. (2003). The Lrp family of transcriptional regulators. Mol Microbiol *48*, 287-294.

Brochier-Armanet, C., and Forterre, P. (2007). Widespread distribution of archaeal reverse gyrase in thermophilic bacteria suggests a complex history of vertical inheritance and lateral gene transfers. Archaea *2*, 83-93.

Brock, T. D., Brock, K. M., Belly, R. T., and Weiss, R. L. (1972). *Sulfolobus*: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. Arch Mikrobiol *84*, 54-68.

Brock, T. D., and Freeze, H. (1969). *Thermus aquaticus* gen. n. and sp. n., a nonsporulating extreme thermophile. J Bacteriol *98*, 289-297.

Brugger, K., Chen, L., Stark, M., Zibat, A., Redder, P., Ruepp, A., Awayez, M., She, Q., Garrett, R. A., and Klenk, H. P. (2007). The genome of *Hyperthermus butylicus*: a sulfur-reducing, peptide fermenting, neutrophilic Crenarchaeote growing up to 108 degrees C. Archaea *2*, 127-135.

Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., Sutton, G. G., Blake, J. A., FitzGerald, L. M., Clayton, R. A., Gocayne, J. D.*, et al.* (1996). Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science *273*, 1058-1073.

Cambillau, C., and Claverie, J. M. (2000). Structural and genomic correlates of hyperthermostability. J Biol Chem *275*, 32383-32386.

Chen, L., Brugger, K., Skovgaard, M., Redder, P., She, Q., Torarinsson, E., Greve, B., Awayez, M., Zibat, A., Klenk, H. P., and Garrett, R. A. (2005). The Genome of *Sulfolobus acidocaldarius*, a Model Organism of the Crenarchaeota. J Bacteriol *187*, 4992-4999.

Cohen, G. N., Barbe, V., Flament, D., Galperin, M., Heilig, R., Lecompte, O., Poch, O., Prieur, D., Querellou, J., Ripp, R.*, et al.* (2003). An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*. Mol Microbiol *47*, 1495-1512.

Daniel, R. M., Danson, M. J., Eisenthal, R., Lee, C. K., and Peterson, M. E. (2008). The effect of temperature on enzyme activity: new insights and their implications. Extremophiles *12*, 51-59.

Deckert, G., Warren, P. V., Gaasterland, T., Young, W. G., Lenox, A. L., Graham, D. E., Overbeek, R., Snead, M. A., Keller, M., Aujay, M.*, et al.* (1998). The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. Nature *392*, 353-358.

Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics *23*, 673-679.

DiDonato, M., Deacon, A. M., Klock, H. E., McMullan, D., and Lesley, S. A. (2004). A scaleable and integrated crystallization pipeline applied to mining the *Thermotoga maritima* proteome. J Struct Funct Genomics *5*, 133-146.

Eddy, S. R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.

Edgar, R. C. (2007). PILER-CR: fast and accurate identification of CRISPR repeats. BMC Bioinformatics *8*, 18.

Eijsink, V. G., Gaseidnes, S., Borchert, T. V., and van den Burg, B. (2005). Directed evolution of enzyme stability. Biomol Eng *22*, 21-30.

Ettema, T. J., de Vos, W. M., and van der Oost, J. (2005). Discovering novel biology by *in silico* archaeology. Nat Rev Microbiol *3*, 859-869.

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

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

Fukui, T., Atomi, H., Kanai, T., Matsumi, R., Fujiwara, S., and Imanaka, T. (2005). Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. Genome Res *15*, 352-363.

Gabaldón, T., and Huynen, M. A. (2004). Prediction of protein function and pathways in the genome era. Cell Mol Life Sci *61*, 930-944.

Galtier, N., and Lobry, J. R. (1997). Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. J Mol Evol *44*, 632-636.

Geiduschek, E. P., and Ouhammouch, M. (2005). Archaeal transcription and its regulators. Mol Microbiol *56*, 1397-1407.

Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S. R., and Bateman, A. (2005). Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res *33*, D121-124.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007). CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. Nucleic Acids Res *35*, W52-57.

Henne, A., Bruggemann, H., Raasch, C., Wiezer, A., Hartsch, T., Liesegang, H., Johann, A., Lienard, T., Gohl, O., Martinez-Arias, R.*, et al.* (2004). The genome sequence of the extreme thermophile *Thermus thermophilus*. Nat Biotechnol *22*, 547-553.

Huang, S. L., Wu, L. C., Liang, H. K., Pan, K. T., Horng, J. T., and Ko, M. T. (2004). PGTdb: a database providing growth temperatures of prokaryotes. Bioinformatics *20*, 276-278.

Kashefi, K., and Lovley, D. R. (2003). Extending the upper temperature limit for life. Science *301*, 934.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A.*, et al.* (2001). Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7. DNA Res *8*, 123-140.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A.*, et al.* (1999). Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1. DNA Res *6*, 83-101, 145-152.

Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A.*, et al.* (1998). Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3. DNA Res *5*, 55-76.

Kikuchi, A., and Asai, K. (1984). Reverse gyrase--a topoisomerase which introduces positive superhelical turns into DNA. Nature *309*, 677-681.

Klein, R. J., Misulovin, Z., and Eddy, S. R. (2002). Noncoding RNA genes identified in AT-rich hyperthermophiles. Proc Natl Acad Sci U S A *99*, 7542-7547.

Klenk, H. P., Clayton, R. A., Tomb, J. F., White, O., Nelson, K. E., Ketchum, K. A., Dodson, R. J., Gwinn, M., Hickey, E. K., Peterson, J. D.*, et al.* (1997). The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon Archaeoglobus fulgidus. Nature *390*, 364-370.

Koonin, E. V. (2001). Computational genomics. Curr Biol *11*, R155-158.

Liolios, K., Mavromatis, K., Tavernarakis, N., and Kyrpides, N. C. (2007). The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res.

Lowe, T. M., and Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res *25*, 955-964.

Lowe, T. M., and Eddy, S. R. (1999). A computational screen for methylation guide snoRNAs in yeast. Science *283*, 1168-1171.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Yadhukumar, Buchner, A., Lai, T., Steppi, S., Jobb, G.*, et al.* (2004). ARB: a software environment for sequence data. Nucleic Acids Res *32*, 1363-1371.

Nakagawa, S., Takaki, Y., Shimamura, S., Reysenbach, A. L., Takai, K., and Horikoshi, K. (2007). Deep-sea vent ε-proteobacterial genomes provide insights into emergence of pathogens. Proc Natl Acad Sci U S A *104*, 12146-12150.

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A.*, et al.* (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature *399*, 323-329.

Nguyen, T. N., Ejaz, A. D., Brancieri, M. A., Mikula, A. M., Nelson, K. E., Gill, S. R., and Noll, K. M. (2004). Whole-genome expression profiling of *Thermotoga maritima* in response to growth on sugars in a chemostat. J Bacteriol *186*, 4824-4828.

Pearson, W. R. (1990). Rapid and sensitive sequence comparison with FASTP and FASTA. Methods Enzymol *183*, 63-98.

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

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., Mullis, K. B., and Erlich, H. A. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science *239*, 487-491.

Saitou, N., and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol *4*, 406-425.

Sato, T., Fukui, T., Atomi, H., and Imanaka, T. (2003). Targeted gene disruption by homologous recombination in the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1. J Bacteriol *185*, 210-220.

Schonheit, P., and Schafer, T. (1995). Metabolism of Hyperthermophiles. World Journal of Microbiology & Biotechnology *11*, 26-57.

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

She, Q., Singh, R. K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M. J., Chan-Weiher, C. C., Clausen, I. G., Curtis, B. A., De Moors, A.*, et al.* (2001). The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc Natl Acad Sci U S A *98*, 7835-7840.

Siebers, B., and Schonheit, P. (2005). Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. Curr Opin Microbiol *8*, 695-705.

Slesarev, A. I., Mezhevaya, K. V., Makarova, K. S., Polushin, N. N., Shcherbinina, O. V., Shakhova, V. V., Belova, G. I., Aravind, L., Natale, D. A., Rogozin, I. B.*, et al.* (2002). The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens. Proc Natl Acad Sci U S A *99*, 4644-4649.

Smith, T. F., and Waterman, M. S. (1981). Identification of common molecular subsequences. J Mol Biol *147*, 195-197.

Stetter, K. O. (1996). Hyperthermophilic procaryotes. FEMS Microbiology Reviews *18*, 149-158.

Thompson, W., Rouchka, E. C., and Lawrence, C. E. (2003). Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res *31*, 3580-3585.

Unsworth, L. D., van der Oost, J., and Koutsopoulos, S. (2007). Hyperthermophilic enzymes--stability, activity and implementation strategies for high temperature applications. Febs J *274*, 4044-4056.

Verhees, C. H., Kengen, S. W., Tuininga, J. E., Schut, G. J., Adams, M. W., De Vos, W. M., and Van Der Oost, J. (2003). The unique features of glycolytic pathways in Archaea. Biochem J *375*, 231-246.

Vieille, C., and Zeikus, G. J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev *65*, 1-43.

von Mering, C., Jensen, L. J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B., and Bork, P. (2007). STRING 7--recent developments in the integration and prediction of protein interactions. Nucleic Acids Res *35*, D358-362.

Voordeckers, J. W., Starovoytov, V., and Vetriani, C. (2005). *Caminibacter mediatlanticus* sp. nov., a thermophilic, chemolithoautotrophic, nitrate-ammonifying bacterium isolated from a deep-sea hydrothermal vent on the Mid-Atlantic Ridge. Int J Syst Evol Microbiol *55*, 773-779.

Waters, E., Hohn, M. J., Ahel, I., Graham, D. E., Adams, M. D., Barnstead, M., Beeson, K. Y., Bibbs, L., Bolanos, R., Keller, M.*, et al.* (2003). The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism. Proc Natl Acad Sci U S A *100*, 12984-12988.

Weinberg, M. V., Schut, G. J., Brehm, S., Datta, S., and Adams, M. W. (2005). Cold shock of a hyperthermophilic archaeon: *Pyrococcus furiosus* exhibits multiple responses to a suboptimal growth temperature with a key role for membrane-bound glycoproteins. J Bacteriol *187*, 336-348.

Woese, C. R., and Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A *74*, 5088-5090.

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A *87*, 4576-4579.

Wu, M., Ren, Q., Durkin, A. S., Daugherty, S. C., Brinkac, L. M., Dodson, R. J., Madupu, R., Sullivan, S. A., Kolonay, J. F., Haft, D. H.*, et al.* (2005). Life in hot carbon monoxide: the complete genome sequence of *Carboxydothermus hydrogenoformans* Z-2901. PLoS Genet *1*, e65.

Xue, Y., Xu, Y., Liu, Y., Ma, Y., and Zhou, P. (2001). *Thermoanaerobacter tengcongensis* sp. nov., a novel anaerobic, saccharolytic, thermophilic bacterium isolated from a hot spring in Tengcong, China. Int J Syst Evol Microbiol *51*, 1335-1341.

Yokoyama, S., Matsuo, Y., Hirota, H., Kigawa, T., Shirouzu, M., Kuroda, Y., Kurumizaka, H., Kawaguchi, S., Ito, Y., Shibata, T.*, et al.* (2000). Structural genomics projects in Japan. Prog Biophys Mol Biol *73*, 363-376.

Zhu, W., Reich, C. I., Olsen, G. J., Giometti, C. S., and Yates, J. R., 3rd (2004). Shotgun proteomics of *Methanococcus jannaschii* and insights into methanogenesis. J Proteome Res *3*, 538-548.

Chapter 2

Reconstruction of central carbon metabolism in *Sulfolobus solfataricus* using a two-dimensional gel electrophoresis map, stable isotope labeling and DNA microarray analysis

Snijders, A. P. L., Walther, J., Peter, S., Kinnman, I., de Vos, M. G. J., **van de Werken, H. J. G.**, Brouns, S. J. J., van der Oost, J., and Wright, P. C. (2006). Proteomics *6*, 1518-1529.

Abstract

In the last decade, an increasing number of sequenced archaeal genomes have become available, opening up the possibility for functional genomic analyses. Here, we reconstructed the central carbon metabolism in the hyperthermophilic crenarchaeon *Sulfolobus solfataricus* (glycolysis, gluconeogenesis and tricarboxylic acid cycle) on the basis of genomic, proteomic, transcriptomic and biochemical data. A 2-DE reference map of *S. solfataricus* grown on glucose, consisting of 325 unique ORFs in 255 protein spots, was created to facilitate this study. The map was then used for a differential expression study based on 15N metabolic labeling (yeast extract 1 tryptone grown cells (YT) *vs.* glucose-grown cells (G)). In addition, the expression ratio of the genes involved in carbon metabolism was studied using DNA microarrays. Surprisingly, only 3 and 14% of the genes and proteins, respectively, involved in central carbon metabolism showed a greater than two-fold change in expression level. All results are discussed in the light of the current understanding of central carbon metabolism in *S. solfataricus* and will help to obtain a system-wide understanding of this organism.

Introduction

Sulfolobus solfataricus is a thermoacidophilic crenarchaeon that grows between 70 and 90 ºC and in a pH range of 2-4 (Zillig *et al.*, 1980). Its preference for environments hostile to many other organisms makes it an interesting source for novel, thermostable enzymes. *S. solfataricus* has been an attractive crenarchaeal model organism since its isolation in the early 1980s, and the completion of the genomic sequence in 2001 (She *et al.*, 2001) has further increased its popularity. Currently, 1941 genes (53.11%) in TIGR's comprehensive microbial resource (CMR) database have no known function (Peterson *et al.*, 2001). Of the 2977 Open Reading Frames (ORFs) originally identified in the genome of *S. solfataricus*, 40% of the genes are archaea specific, 12% are bacteria specific and 2.3% are shared exclusively with eukaryotes. Currently, genetic tools are under development that will contribute to our understanding of fundamental processes in *Sulfolobus* (Contursi *et al.*, 2003; Jonuscheit *et al.*, 2003; Limauro *et al.*, 2001; Stedman *et al.*, 1999; Worthington *et al.*, 2003). In order to fully exploit its potential for metabolic engineering, a deeper understanding of the central energy and precursor generating pathways is necessary.

The central metabolic pathways in archaea contain many unique features compared to the classical pathways in bacteria and eukaryotes (Adams *et al.*, 2001; Verhees *et al.*, 2003). In *S. solfataricus*, glucose degradation proceeds via a non-phosphorylated version of the Entner-Doudoroff (ED) pathway (De Rosa *et al.*, 1984; Schafer, 1996; Schonheit and Schafer, 1995). In this pathway, glucose is converted into pyruvate through the action of glucose dehydrogenase, gluconate dehydratase, 2-keto-3-deoxy-gluconate (KDG) aldolase, glyceraldehyde dehydrogenase, glycerate kinase, enolase and pyruvate kinase. Recently, experimental evidence has been provided for the operation of the semi-phosphorylated ED pathway in *S. solfataricus* in which KDG is phosphorylated (Ahmed *et al.*, 2005). Gluconeogenesis via a reversed ED pathway is unlikely, since the key enzymes in this pathway do not seem to be able to distinguish between glucose and galactose derivatives. In this case, gluconeogenesis via a reversed ED pathway would result in a mixture of glucose and galactose (Lamble *et al.*, 2003). Instead, *in silico* analysis of the *Sulfolobus* genomes as well as experimental evidence has revealed the presence of a near complete set of proteins involved in the Embden-Meyerhof-Parnas (EMP) pathway (Verhees *et al.*, 2003), suggested to be active in the gluconeogenic direction rather than in the glycolytic direction (Lamble *et al.*, 2003).

In this study, we reconstructed central carbon metabolism and the TCA cycle on the basis of biochemical, computational, proteomic and DNA microarray data, obtained from cell extracts of *S. solfataricus* grown on sugars and peptides. First of all, a two-Dimensional gel Electrophoresis (2-DE) map was created to provide a global overview of protein expression under glucose degrading conditions. This map was then used to investigate the relative abundance of proteins involved in sugar metabolism under minimal or rich media through a ^{15}N metabolic labeling approach. Moreover, DNA microarray analysis was performed to compare mRNA expression under the same conditions. In the last few years, similar transcriptome studies have been conducted with several archaea that utilize different types of glycolysis. These organisms include: *Pyrococcus furiosus* (Schut *et al.*, 2003) an obligate anaerobic hyperthermophile with an EMP-like pathway and *Haloferax volcanii* (Zaigler *et al.*, 2003) a facultative anaerobic halophile using an ED-like glycolysis. However, there are relatively few studies that combine transcriptomics and proteomics, and none have so far been published for archaea.

Here, we present a study in which both quantitative proteomics and transcriptomics were used to analyze the expression of the genes involved in the central carbon metabolism of *Sulfolobus solfataricus*

Materials and methods

Cell growth and harvest

Sulfolobus solfataricus P2 (DSM1617) was grown aerobically in a rotary shaker at 80 ºC in a medium of pH 3.5-4.0 which contained: 2.5 g/L $(NH_4)_2SO_4$, 3.1 g/L KH_2PO_4, 203.3 mg/L $MgCl_2$ • 6 H_2O, 70.8 mg/L $Ca(NO_3)_2$ • 4 H_2O, 2 mg/L $FeSO_4$ • 7 H_2O, 1.8 mg/L $MnCl_2$ • 4 H_2O, 4.5 mg/L $Na_2B_4O_7$ • 2 H_2O, 0.22 mg/L $ZnSO_4$ • 7 H_2O, 0.06 mg/L $CuCl_2$ • 2 H_2O, 0.03 mg/L Na_2MoO_4 • 2 H_2O, 0.03 mg/L $VOSO_4$ • 2 H_2O, 0.01 mg/L $CoCl_2$ • 6 H_2O. The medium was supplemented with Wollin vitamins, and either 0.3% to 0.4% D-glucose (G) or 0.1% Yeast extract and 0.2% Tryptone (YT). The Wollin vitamin stock (100x) contained 2 mg/L D-Biotin, 2 mg/L Folic acid, 10 mg/L Pyridoxine-HCl, 10 mg/L Riboflavin, 5 mg/L Thiamine-HCl, 5 mg/L Nicotinic acid, 5 mg/L DL-Ca-Pantothenate, 0.1 mg/L Vitamin B12, 5 mg/L p-Aminobenzoic acid, 5 mg/L Lipoic acid. Cell growth was monitored by measuring the turbidity at 530 or 600 nm. Cells for the proteome reference map were harvested by centrifugation in the late exponential growth phase at an OD_{530} of 1.0. Cells were washed twice with a 10 mM Tris-HCl Buffer (pH 7). Subsequently, cells were stored at –20ºC until required. During this whole process, considerable care was taken to ensure that culture to culture variation was minimized, and cultures were prepared in at least triplicate. In the case of the ^{15}N labeling experiment, $(^{15}NH4)2SO4$ was used as the nitrogen source. Cells were incubated with ^{15}N ammonium sulfate for at least 8 doubling times to allow for full incorporation of the label. After this, the ^{14}N and ^{15}N growth experiments were set up simultaneously. When the optical density reached a value of 0.5, the cultures were mixed. To ensure that equal amounts of biomass were mixed, slight corrections in volume were made in case the OD_{530} was not exactly 0.5. Previously, we have demonstrated that this approach leads to accurate mixing (Snijders *et al.*, 2005b). Next, cells were pelleted by centrifugation, washed twice with a 10 mM Tris-HCl Buffer (pH 7) and stored at -20°C. Preparation of cell extracts, 2-DE and protein identification was performed in exactly the same manner for the labeled/unlabeled cells as for the unlabeled cells.

Preparation of cell extracts

The -20°C frozen cells were thawed and immediately resuspended in 1.5 ml of 10 mM Tris-HCl buffer (pH 7), and 25 µl of a protease-inhibitor cocktail (Sigma) was added. Cells were disrupted by sonication for 10 minutes on ice ("Soniprep 150", Sanyo). Insoluble cell material was removed by centrifugation at 13,000 rpm for 10 minutes. The protein concentration of the supernatant was determined using the Bradford Protein Assay (Sigma). The supernatant was subsequently stored at –80°C.

Two-dimensional gel electrophoresis (2-DE)

Gels for the reference map were prepared in triplicate. The extract was mixed with a rehydration buffer containing 50 mM DTT (Sigma), 8 M Urea (Sigma), 2% CHAPS (Sigma), 0.2% (w/v) Pharmalyte ampholytes pH 3-10 (Fluka) and Bromophenol Blue (trace) (Sigma). This mixture was designated as the sample mix. Three IPG strips (pH 3-10) (Bio-Rad) were rehydrated with 300 µl (400 µg) of this sample mix. Strips were allowed to rehydrate overnight. IEF was performed using a 3-step protocol at a temperature of 20°C using a Protean IEF cell (Bio-Rad). In the first step, the voltage was linearly ramped to 250 V over 30 minutes to desalt the strips. Next, the voltage was linearly ramped to 1000 V over 2.5 half hours. Finally, the voltage was rapidly ramped to 10,000 V for 40,000 V/hours to complete the focusing. At this stage, the strips were stored overnight at –20°C. Focused strips were first incubated for 15 minutes in a solution containing 6M Urea, 2% SDS, 0.375 M Tris-HCl (pH 8.8), 20% glycerol, and 2% (w/v) DTT. After this, the solution was discarded and the strips were incubated in a solution containing 6 M Urea, 2% SDS, 0.375 M Tris-HCl (pH 8.8), 20% Glycerol, and 4% Iodoacetamide. After equilibration, proteins were separated in the second dimension using SDS-PAGE performed using a Protean II Multicell (Bio-Rad) apparatus on 10% T, 2.6% C gels (17 cm x 17 cm x 1 mm). Electrophoresis was carried out with a constant current of 16 mA/gel for 30 minutes; subsequently the current was increased to 24 mA/gel for another 7 hours.

Protein visualization and image analysis

Gels were stained using Coomassie Brilliant Blue G250 (Sigma). Gels were scanned using a GS-800 densitometer (Bio-Rad) at 100 microns resolution. All spot detection and quantification was performed with PDQUEST 7.1.0 (Bio-Rad). Staining intensity was normalized against the total staining intensity on the gel. 255 spots were selected for mass spectrometric analysis. For protein quantitation, metabolic labeling was used, and for this gel image was matched to the reference map and protein spots of interest were selected for MS analysis and quantitation.

Protein isolation and identification by MS

Spots of interest were excised from the stained 2-DE gels by hand, destained with 200 mM ammonium bicarbonate with 40% acetonitrile. The gel pieces were incubated overnight in a 0.4 µg trypsin solution (Sigma) and 50 µl of 40 mM ammonium bicarbonate in 9% acetonitrile. The next day, peptides were extracted in three subsequent extraction steps using 5 µl of 25 mM NH_4HCO_3 (10 minutes, room temperature), 30 µl acetonitrile (15 minutes, 37^{o}C), 50 µl of 5% formic acid (15 minutes, 37^{o}C) and finally with 30 µl acetonitrile (15 minutes, 37^{o}C). All extracts were pooled and dried in a vacuum centrifuge, then stored at –20^{o}C.

The lyophilized peptide mixture was resuspended in 0.1% formic acid in 3% acetonitrile. This mixture was separated on a PepMap C-18 RP capillary column (LC Packings, Amsterdam, the Netherlands) and eluted in a 30-minute gradient via a LC Packings Ultimate nanoLC directly onto the mass spectrometer. Peptides were analyzed using an Applied Biosystems QStarXL® electrospray ionization quadrupole time of flight tandem mass spectrometer (ESI qQ-TOF). The data acquisition on the MS was performed in the positive ion mode using Information Dependent Acquisition (IDA). Peptides with charge states 2 and 3 were selected for tandem mass spectrometry. IDA data were submitted to Mascot for database searching in a sequence query type of search (www.matrixscience.com). The peptide tolerance was set to 2.0 Da and the MS/MS tolerance was set to 0.8 Da. A carbamidomethyl modification of cysteine was set as a fixed modification and methionine oxidation was set as a variable modification. Up to 1 missed cleavage site by trypsin was allowed. The search was performed against the Mass Spectrometry protein sequence DataBase (MSDB; ftp://ftp.ncbi.nih.gov/repository/MSDB/msdb.nam). Molecular Weight Search (MOWSE) (Pappin *et al.*, 1993) scores greater than 50 were regarded as significant.

Peptide quantitation

In the metabolic labeling experiments, peptide identification of the light (^{14}N) version of the peptide was performed as described above. After this the heavy ^{15}N version of the peptide could be identified by changing the isotope abundance of ^{15}N nitrogen to 100% in the Analyst software data dictionary. Next, the peak area of both version of the same peptide was integrated over time using LC-MS reconstruction tool in the Analyst software. In addition, an extracted ion chromatogram (XIC) was constructed for each peptide. The XIC is an ion chromatogram that shows the intensity values of a single mass (peptide) over a range of scans. This tool was used to check for chromatographic shifts between heavy and light versions of the same peptide.

RNA extraction and probe synthesis

Early-log phase cultures (OD_{600} 0.1-0.2) of *S. solfataricus* grown on 0.1% yeast extract and 0.2% tryptone (YT) or 0.3% D-glucose (G) were quickly cooled in ice-water and harvested by centrifugation at 4 ºC. The RNA extraction was done as described previously (Brinkman *et al.*, 2002). Preparation of cDNA was done as follows: to 15 µg of RNA, 5 µg of random hexamers (Qiagen) was added in a total volume of 11.6 µl. This was incubated for ten minutes at 72 ºC after which the mixture was cooled on ice. Next, dATP, dGTP and dCTP (5 µM final concentration) were added, together with 4 µM aminoallyl dUTP (Sigma), 1 µM dTTP, 10 mM dithiotreitol (DTT), 400 U superscript II (Invitrogen) and the corresponding 5x RT buffer in a final volume of 20 µl. The reverse transcriptase reaction was carried out at 42 ºC for one hour. To stop the reaction and to degrade the RNA, 2 µl 200 mM EDTA and 3 µl 1 M NaOH were added to the reaction mixture, after which it was incubated at 70 ºC for 15 minutes. After neutralization by the addition of 3 µl 1 M HCl, the cDNA was purified using a Qiagen MinElute kit according to the manufacturer's instructions, except that the wash buffer was replaced with 80% (v/v) ethanol. The cDNA was then labeled using post labeling reactive CyDye packs (Amersham Biosciences), according to the protocol provided by the company. Differentially labeled cDNA derived from *S. solfataricus* cells grown on either YT or G media was pooled (15 µg labeled cDNA of each sample) and excess label was removed by cDNA purification using the MinElute kit.

DNA microarray hybridization, scanning and data analysis

The design and construction of the microarray, as well as the hybridization was performed as described previously (Andersson *et al.*, 2005; Lundgren *et al.*, 2004). After hybridization, the microarrays were scanned at a resolution of 5 microns with a Genepix 4000B scanner (Axon Instruments) using the appropriate laser and filter settings. Spots were analyzed with the Genepix pro 5.0 software package (Axon Instruments). Low-quality spots were excluded using criteria that were previously described (Lundgren *et al.*, 2004). Log_2 transformed ratios (log_2(YT/G)) from the replicate slides were averaged after first averaging the duplicate spots on the array. Statistical significance for the observed ratios was calculated by doing a Significance Analysis of Microarrays (SAM) analysis (Tusher *et al.*, 2001). Each log_2 value represents 2 hybridization experiments, performed in duplicate by using cDNA derived from four different cultures of *S. solfataricus*: two grown on YT media and two grown on glucose media. The result of each ORF therefore consisted of 8 pairwise comparisons. The ORFs were categorized according to the 20 functional categories of the comprehensive microbial resource (CMR) (Peterson *et al.*, 2001).

Metabolic pathway reconstruction based on biochemical and genomic data

The reconstruction of the main metabolic pathways was performed with BLASTP and PSI-BLAST programs (Altschul *et al.*, 1997) on the non-redundant (NR) database of protein sequences (National Center for Biotechnology Information) by using full-length or N-terminal protein sequences. All the sequences were derived from verified enzymatic activities of thermophilic or hyperthermophilic archaea unless stated otherwise. The sequences from *Sulfolobus acidocaldarius* were analyzed by BLASTP program using the complete genome sequence (Chen *et al.*, 2005). All the assigned enzymatic functions for the proteins of *Sulfolobus solfataricus* P2 were checked with the annotations in public protein databases, such as the BRaunschweig ENzyme DAtabase (BRENDA) (Schomburg *et al.*, 2004), Clusters of Orthologous Groups of proteins (COG) (Tatusov *et al.*, 2003), InterPro (Mulder *et al.*, 2005) and the fee-based ERGO bioinformatics suite (Overbeek *et al.*, 2003). The reconstructed pathways were compared with previous reports (Huynen *et al.*, 1999; Ronimus and Morgan, 2003; Verhees *et al.*, 2003) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.*, 2004).

Results and Discussion

Generation and application of a two-dimensional gel electrophoresis map

Figure 2.1 shows an image of the 2-DE reference map for S. solfataricus. With Coomassie Brilliant blue G250, approximately 500 spots were visualized. The highest spot count was obtained in the region pI = 5-9, and proteins ranged in size from 15-123 kDa (predicted values). In total, 255 spots were selected for Mass Spectrometry (MS) analysis on the basis of their relative high abundance. In addition, faint spots were selected to test the sensitivity of the MS method. In total, 325 unique proteins in 255 spots were identified, with even the faintest spots yielding significant Molecular Weight Search (MOWSE) scores (> 51). All 255 spots were found on the triplicate gels. The complete dataset is presented in the supplementary material. A subset, representing key elements of central energy metabolism and other relevant proteins is discussed more extensively in this paper. The highest MOWSE score, 1362, was achieved for elongation factor 2 (Sso0728, spot 26). Generally, one peptide (intact mass and tandem mass spectrometry (MS/MS) ion spectrum) was sufficient for confident identification of a S. solfataricus protein against the full Mass Spectrometry protein sequence Database (MSDB). In most cases, however, multiple peptides of the same protein were recovered from a spot. On average, the sequence coverage was 30%. The highest sequence coverage (75%) was found for the α-subunit of the proteasome (Sso0738) in spot 213. There was no correlation between the sequence coverage and the protein size. However, larger proteins usually resulted in higher MOWSE scores. This is due to the fact that larger proteins generate a larger number of unique peptides after tryptic digestion. For example, MOWSE scores greater than 800 were only obtained for proteins larger than 48 kDa.

The number of proteins that matched to ORFs that are either hypothetical or conserved hypothetical proteins was 157 (48%). This is similar compared to the expected 53%, on the basis of the genome composition. This was also found in a similar study on the *Methanocaldococcus jannaschii* proteome (Giometti *et al.*, 2002). Interestingly, there were only two hypothetical proteins amongst the 20 most intense spots, (Sso0029, Sso0099 relating to spots 130 and 224 respectively). The relatively high abundance of those proteins suggests an important function.

Another important observation is that a number of proteins were found in more than

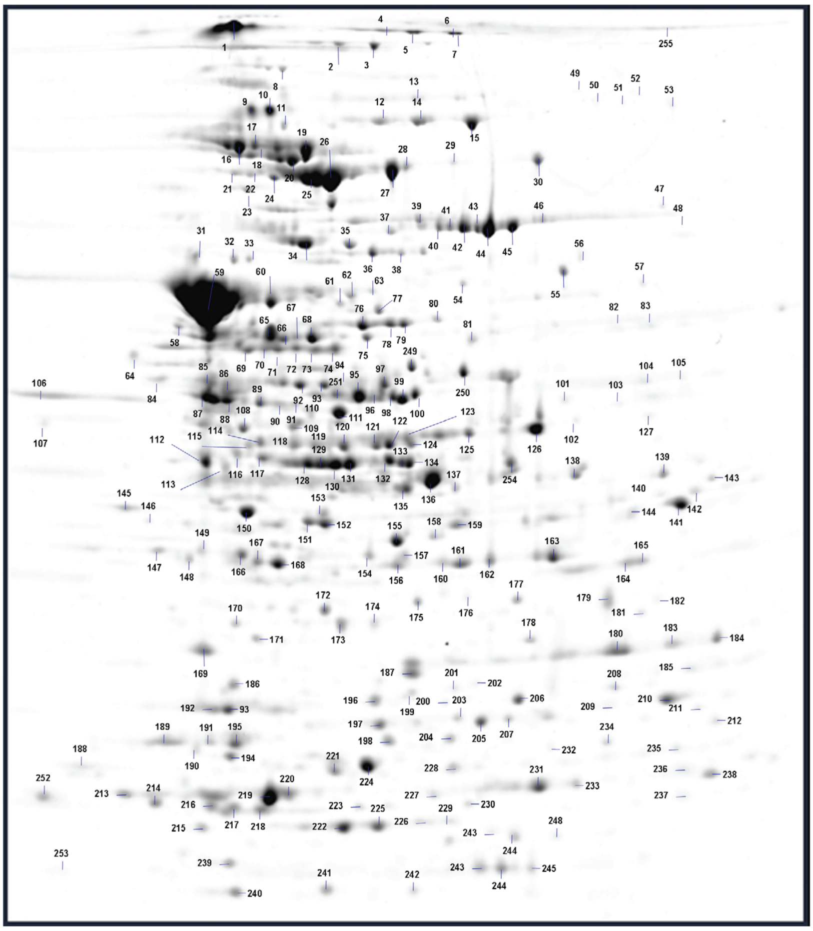


Figure 2.1 2-DE reference map for S. solfataricus grown on glucose. All numbered spots were subjected to LC-MS-MS analysis. Results are displayed in Table 1 (supplementary material).

one spot. Interestingly, this was true for a large number of proteins involved in the TCA cycle (*e.g.* 2-oxoacid:ferredoxin oxidoreductase (Sso2815) was found in eight different spots). There are a number of explanations for this: (1) Isoforms or post-translationally modified versions of the protein might be present in the cell (2) the protein was modified during protein extraction or during 2-DE (e.g. proteolysis, methionine oxidation,), (3) the protein does not resolve well on the gel and therefore "smears" out over a large pH or mass range, or (4) the denaturating conditions are not strong enough to completely break protein associations. The presence of a protein in multiple spots was also observed in similar proteomic studies (Giometti *et al.*, 2002). To find post-translational modifications, all mass spectra were searched again but this time with phosphorylation of serine or threonine, and with methylation set as variable modifications. Unfortunately, no consistent results were obtained, and therefore more specific studies targeted to identify post-translational modifications are necessary.

In a number of cases multiple proteins per spot were found. Often these proteins have similar molecular weights (MW) and iso-electric points (pI) indicating that the resolution on the gel was insufficient to resolve these proteins into single protein spots. In other cases however, proteins in the same spot differ significantly in MW and pI. These represent biologically interesting cases since these could indicate stable protein associations. An example was found in spot 1, where subunits α, β and γ of aldehyde oxidoreductase (Sso2636, Sso2637, Sso2639) were found.

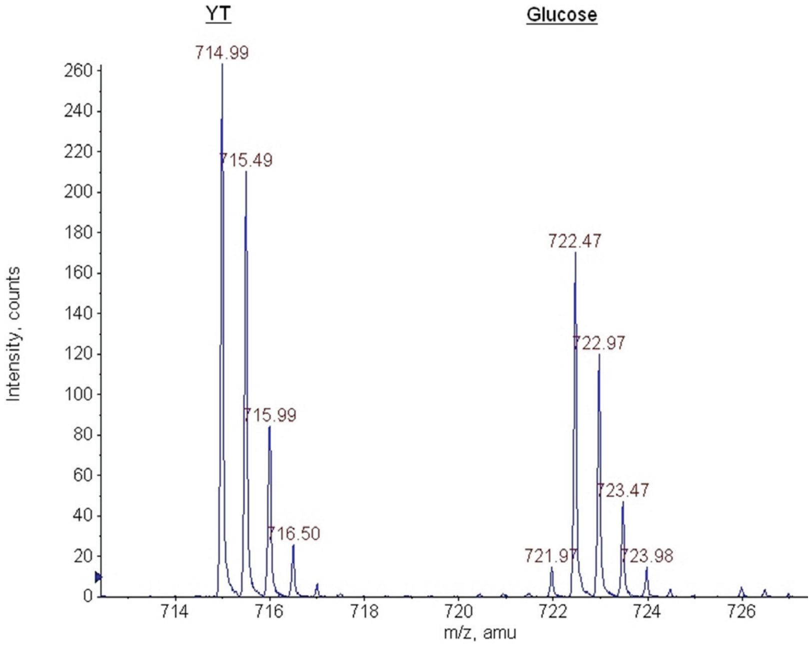


Figure 2.2 Peptide quantitation. TOF MS spectrum of a ^{15}N labeled and an unlabeled peptide. The peak on the left at m/z 714.99 represents the unlabeled version of the peptide (protein from cells grown on yeast extract + tryptone (YT)). The peak at the right at m/z 7.22.47 represents the ^{15}N labeled version of the peptide (protein from cells grown on glucose). This peptide was identified as IFGSLSSNYVLTK, corresponding to 2-keto-3-deoxy-gluconate aldolase (Sso3197). The ratio between the areas of the heavy and light versions of this peptide was 1.56.

Protein quantitation was performed on the basis of ^{15}N metabolic labeling as recently described. With this method a number of problems associated with 2-DE (e.g. multiple proteins per spot) can be avoided. Moreover, the reproducibility of gel staining becomes of lesser importance since protein quantitation takes place on the MS (Snijders *et al.*, 2005a).

Figure 2.2 shows an example of a TOF-MS spectrum containing both the light and the heavy version of the peptide IFGSLSSNYVLTK. This peptide is derived from the 2-keto-3-deoxy-gluconate aldolase (Sso3197). The light peptide at m/z 714.99 corresponds to the yeast extract + tryptone (YT) grown cells and the heavy peptide at m/z 722.47 corresponds to the glucose (G) grown cells. The relative abundance of the heavy and light peptide can now be calculated by determining the ratio of the peak areas. Note that the difference in mass between the heavy and light version of the peptide corresponded exactly to the number of nitrogen atoms in the peptide, in this case 15 atoms ($\Delta M/z = 7.5$). Table 2.1 summarizes the differential proteomic data obtained in this way, as well as the corresponding transcriptomic data.

Exploration of the transcriptome

In total, 1581 of the 2315 genes printed on the microarray were used in the analysis (selected, according to criteria described above). There were 184 significantly differentially expressed genes ($p<0.05$; p is the statistical certainty that the observed change in ratio is not caused by a biological effect). In total, 135 and 49 genes are up-regulated under glucose and YT conditions, respectively. Of these up-regulated genes 23% and 20% were annotated as either hypothetical or conserved hypothetical. Interestingly, these percentages are lower than the expected 53%.

Genes involved in amino acid biosynthesis were regulated under both glucose and YT conditions. This was 16% and 10%, of the total amount of upregulated genes, in the case of glucose and YT, respectively. Regulation in this functional group was expected since amino acids are synthesized under glucose conditions and predominantly degraded under YT conditions. This data, therefore, provides an excellent starting point for amino acid metabolism reconstruction. Future biochemical and proteomic studies are necessary to confirm the exact composition and direction of the responsible pathways.

In addition, three genes involved in nitrogen metabolism were regulated: (1) glutamate synthase (Sso0684, 0.15) (2) glutamine synthase (Sso0366; 0.27) and (3) glutamate dehydrogenase (Sso2044; 6.29), absolute ratios are given as YT/G). These results show that cells which grow on glucose assimilate nitrogen by the sequential action of glutamine synthase and glutamate synthase. Under YT conditions glutamate dehydrogenase produces free ammonium by converting glutamate into 2-oxoglutarate. This is necessary because there is an excess of nitrogen bound to carbon when grown in the presence of YT.

Transport and binding proteins are also a major group of up-regulated genes (12% and 8% for glucose and YT, respectively). Previously, it was shown that both glucose and YT grown

Table 2.1 Relative abundances of mRNA and protein levels of the genes involved in central metabolic pathways of *Sulfolobus solfataricus* grown on yeast extract and tryptone (YT) compared to glucose (G).

Locus	Enzyme description	EC	COG	Transcriptomics[a]	Proteomics[a]	Reference
Glycolysis						
Sso3003	Glucose-1-dehydrogenase	1.1.1.47	1063	NS	NF	(Lamble *et al.*, 2003)
Sso2705	Gluconolactonase	3.1.1.17	3386	1.15 ± 0.07	NF	(Verhees *et al.*, 2003)
Sso3041	Gluconolactonase	3.1.1.17	3386	NF	NF	
Sso3198	Gluconate dehydratase	4.2.1.39	4948	1.00 ± 0.07	1.42 ± 0.14	(Kim and Lee, 2005; Lamble *et al.*, 2004)
Sso3197	2-keto-3-deoxy-gluconate aldolase	4.1.2.-	0329	0.96 ± 0.19	1.55 ± 0.05	(Buchanan *et al.*, 1999)
Sso3195	2-keto-3-deoxy-gluconate kinase	2.7.1.45	0524	1.19 ± 0.15	NF	(Verhees *et al.*, 2003)
Sso3194	Glyceraldehyde-3-phosphate dehydrogenase (non-phosphorylating)	1.2.1.3	1012	0.87 ± 0.10	0.66 ± 0.07	(Ahmed *et al.*, 2005; Brunner *et al.*, 1998)
Sso2639[c]	Aldehyde oxidoreductase, α-subunit	1.2.7.-	1529	0.65 ± 0.01[b]	4.51 ± 0.78	(Kardinahl *et al.*, 1999)
Sso2636[c]	Aldehyde oxidoreductase, β-subunit		1319	0.55 ± 0.13[b]	4.89 ± 0.40	
Sso2637[c]	Aldehyde oxidoreductase, γ-subunit		2080	0.62 ± 0.14[b]	4.22 ± 1.03	
Sso0666	Glycerate kinase	2.7.1.-	2379	0.70 ± 0.24	NF	(De Rosa *et al.*, 1984; Verhees *et al.*, 2003)
Sso0981	Pyruvate kinase	2.7.1.40	0469	0.98 ± 0.10	NF	(Schramm *et al.*, 2000)
Glycolysis/Gluconeogenesis						
Sso0417	Phosphoglycerate mutase	5.4.2.1	3635	1.03 ± 0.13	1.55 ± 0.14	(Van der Oost *et al.*, 2002)
Sso2236	Phosphoglycerate mutase	5.4.2.1	0406	NS	NF	
Sso0913	Enolase	4.2.1.11	0148	1.36 ± 0.35	1.59 ± 0.23	(Peak *et al.*, 1994)
Gluconeogenesis						
Sso0883	Phospho*enol*pyruvate synthase	2.7.9.2	0574	1.62 ± 0.08[b]	1.77 ± 0.22	(Hutchins *et al.*, 2001)
Sso0527	Phosphoglycerate kinase	2.7.2.3	0126	1.26 ± 0.26	2.30 ± 0.28	(Hess *et al.*, 1995)
Sso0528	Glyceraldehyde-3-phosphate dehydrogenase (phosphorylating)	1.2.1.12	0057	1.07 ± 0.20	1.16 ± 0.02	(Russo *et al.*, 1995)
Sso2592	Triose-phosphate isomerase	5.3.1.1	0149	NF	1.17 ± 0.12	(Kohlhoff *et al.*, 1996)
Sso3226	Fructose-bisphosphate aldolase	4.1.2.13	1830	NS	1.84 ± 0.10	(Siebers *et al.*, 2001)
Sso0286	Fructose-bisphosphatase	3.1.3.11	1980	1.24 ± 0.18	1.32 ± 0.05	(Nishimasu *et al.*, 2004)
Sso2281	Glucose-6-phosphate isomerase	5.3.1.9	0166	1.01 ± 0.13	1.51 ± 0.10	(Hansen *et al.*, 2004)
Sso0207	Phosphoglucomutase	5.4.2.2	1109	1.03 ± 0.32	1.55 ± 0.01	(Solow *et al.*, 1998)
Tricarboxylic acid cycle						
Sso2589	Citrate synthase	2.3.3.1	0372	0.84 ± 0.09	1.02 ± 0.03	(Lohlein-Werhahn *et al.*, 1988; Smith *et al.*, 1987)

Sso1095	Aconitase	4.2.1.3	1048	1.05 ± 0.14	1.11 ± 0.03	(Uhrigshardt *et al.*, 2001)
Sso2182	Isocitrate dehydrogenase	1.1.1.42	0538	1.34 ± 0.65	1.18 ± 0.03	(Camacho *et al.*, 1995)
Sso2815[d]	2-oxoacid:ferredoxin oxidoreductase α/γ-subunit	1.2.7.1 1.2.7.3	0674 1014	0.89 ± 0.07	0.56 ± 0.05	(Fukuda and Wakagi, 2002; Kerscher *et al.*, 1982; Zhang *et al.*, 1996)
Sso2816[d]	2-oxoacid:ferredoxin oxidoreductase β-subunit		1013	0.85 ± 0.31	0.60 ± 0.02	
Sso2482	Succinate-CoA ligase, α-subunit	6.2.1.5	0074	0.93 ± 0.25	0.54 ± 0.04	(Danson *et al.*, 1985)
Sso2483	Succinate-CoA ligase, β-subunit		0045	0.94 ± 0.30	0.51± 0.05	
Sso2356	Succinate dehydrogenase, subunit A	1.3.99.1	1053	NS	0.58 ± 0.4	(Janssen *et al.*, 1997)
Sso2357	Succinate dehydrogenase, subunit B		0479	0.75 ± 0.28	NF	
Sso2358	Succinate dehydrogenase, subunit C		2048	0.94 ± 0.27	NF	
Sso2359	Succinate dehydrogenase, subunit D			0.89 ± 0.16	NF	
Sso1077	Fumarate hydratase	4.2.1.2	0114	1.08 ± 0.10	1.53 ± 0.09	(Colombo *et al.*, 1994; Puchegger *et al.*, 1990)
Sso2585	Malate dehydrogenase	1.1.1.37	0039	0.82 ± 0.27	0.69 ± 0.01	(Hartl *et al.*, 1987)
Glyoxylate shunt						
Sso1333	Isocitrate lyase	4.1.3.1	2224	0.30 ± 0.07[b]	NF	(Uhrigshardt *et al.*, 2002)
Sso1334	Malate synthase	2.3.3.9	2225	1.11 ± 0.47	1.18 ± 0.04	
C3/C4 interconversions						
Sso2869	Malic enzyme	1.1.1.38	0281	1.05 ± 0.24	1.92 ± 0.15	(Bartolucci *et al.*, 1987)
Sso2537	Phosphoenolpyruvate carboxykinase	4.1.1.32	1274	1.42 ± 0.42	NF	(Fukuda *et al.*, 2004)
Sso2256	Phosphoenolpyruvate carboxylase	4.1.1.31	1892	0.83 ± 0.18	0.88 ± 0.17	(Ettema *et al.*, 2004; Sako *et al.*, 1996)

NF: not found, NS: no significant signal.

[a] relative abundance ratio with standard deviation Yeast extract + Tryptone grown cells / Glucose grown cells (YT/G)

[b] Probability value (p) smaller than 0.05.

[c] enzyme complex has broad substrate specificity for aldehydes

[d] exhibits pyruvate, 2-oxoglutarate and 2-oxobutyrate oxidoreductase activity

cells have the capacity to transport glucose (Elferink *et al.*, 2001). This is reflected by the fact that the genes involved in glucose transport were not differentially expressed (Sso2847, Sso2848, Sso2849, Sso2850). In addition, genes involved in dipeptide transport were up-regulated under YT conditions (Sso1282; 2.01 / Sso2615; 1.74 / Sso2616; 1.57). Interestingly, genes involved in maltose transport were slightly up-regulated under glucose conditions (Sso3053; 0.36 / Sso3058; 0.50 / Sso3059; 0.53).

Metabolic pathway reconstruction

During the last two decades, the main metabolic pathways in *Sulfolobus* spp. have been the subject of extensive experimental research. This has led to a profound understanding of the enzymes and protein complexes that are involved in the glycolysis, the tricarboxylic acid cycle (TCA) and related metabolic conversions (Danson, 1988; Verhees *et al.*, 2003). The availability of the genome sequences of *S. solfataricus* (She *et al.*, 2001), *S. tokodaii* (Kawarabayasi *et al.*, 2001) and *S. acidocaldarius* (Chen *et al.*, 2005) has recently allowed for the identification of the genes encoding these proteins by matching full-length or N-terminal protein sequences to the predicted proteomes. A reconstruction of the central carbon metabolic pathways in *Sulfolobus solfataricus* was performed (Fig. 2.3). The results should be taken with a degree of caution since significant differences exist in the physiology between the three *Sulfolobus* species (Schafer, 1996). Almost all proteins involved in this scheme have been experimentally verified in either *Sulfolobus* spp. or other hyperthermophilic Archaea, such as *Thermoproteus tenax*, *Archaeoglobus fulgidus*, *Thermoplasma acidophilum*, *Pyrococcus furiosus*, *Thermococcus kodakaraensis*, *Methanothermus fervidus* and *Methanocaldococcus jannaschii*. Moreover, the vast majority of the anticipated proteins in *Sulfolobus solfataricus* were found on the 2-DE reference map (Fig. 2.3). On average, the TCA cycle proteins made up approximately 12% of the total staining intensity.

Glycolysis and gluconeogenesis

The genus *Sulfolobus* is known to degrade glucose according to a modified version of the Entner-Doudoroff (ED) pathway. While in most cases phosphorylation in the bacterial ED pathway occurs at the level of glucose, gluconate or 2-keto-3-deoxygluconate (KDG), *S. solfataricus* has been reported to utilize a non-phosphorylated version of the ED pathway, which phosphorylates only at the level of glycerate (De Rosa *et al.*, 1984; Selig *et al.*, 1997). Recent experimental findings, however, indicated the presence of a semi-phosphorylated ED pathway, in which KDG is phosphorylated and subsequently cleaved forming pyruvate and glyceraldehyde-3-phosphate (GAP) by the action of the KDG kinase (Sso3195) and the KDG aldolase (Sso3197) respectively. GAP is then oxidized by a non-phosphorylating GAP dehydrogenase (GAPN, Sso3194) forming 3-phosphoglycerate (3PG) (Ahmed *et al.*, 2005). The only net difference

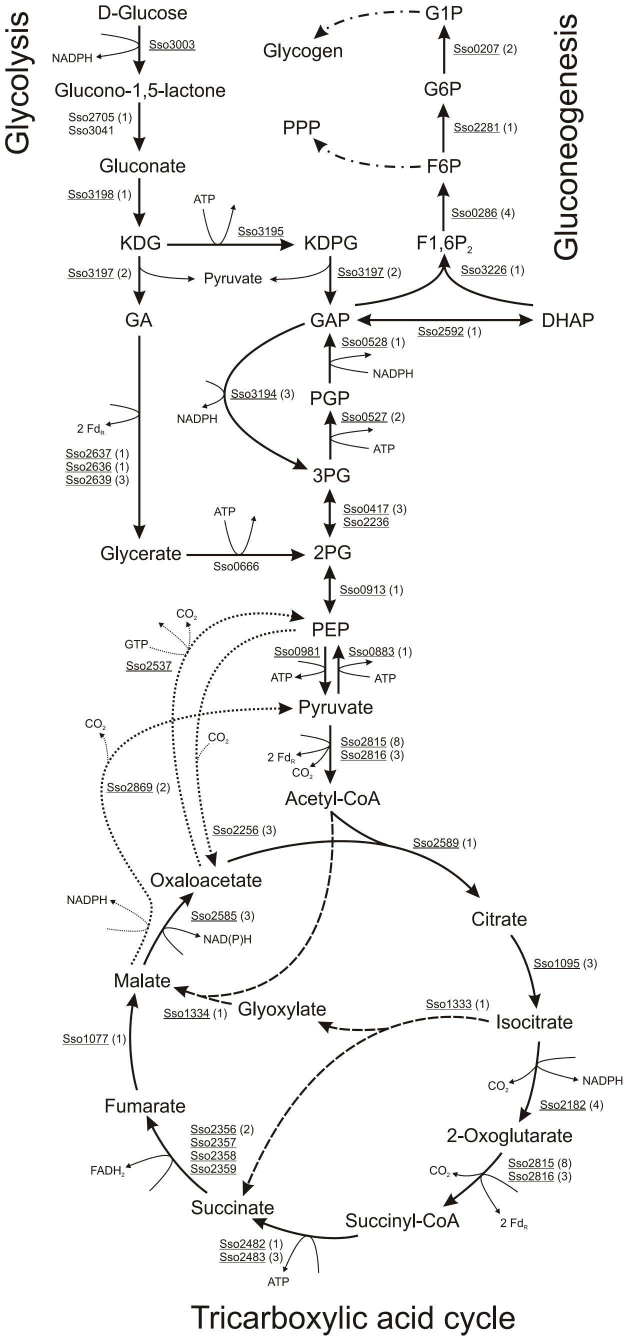


Figure 2.3 Reconstruction of the central metabolic pathways in Sulfolobus solfataricus. Genes involved in the glycolysis, gluconeogenesis and citric acid cycle were surveyed and are indicated by their locus name. Underlined genes were experimentally verified in *Sulfolobus* or related hyperthermophilic Archaea (Table 2.1). The number of spots that were found on the 2-DE reference map is indicated between brackets. The glyoxylate shunt in shown by dashed arrows, while the three to four carbon interconversions are depicted by dotted arrows. Mixed dashed and dotted arrows indicate that the exact pathway to glycogen and pentoses is unknown. The following abbreviations were used: KD(P)G 2-keto-3-deoxy-D-gluconate-(6-phosphate), GA(P) glyceraldehyde-(3-phosphate), PGP 1,3-bisphosphoglycerate, 3PG 3-phosphoglycerate, 2PG 2-phosphoglycerate, PEP phospho*enol*pyruvate, DHAP dihydroxyacetonephosphate, F1,6P$_2$ fructose-1,6-bisphosphate, F6P fructose-6-phosphate, G6P glucose-6-phosphate, G1P glucose-1-phosphate, Fd$_R$ reduced ferredoxin, PPP pentose phosphate pathway. NAD(P)H indicates that both NAD$^+$ and NADP$^+$ can be used as a cofactor. Arrows represent the presumed physiologically relevant direction of catalysis and are not indicative of enzymatic reversibility.

between the non- and semi-phosphorylated pathways is the fact that either reduced ferredoxin (Fd_R) or NADPH is produced, since neither pathway directly yields ATP by substrate level phosphorylation.

The intrinsic irreversibility of several ED enzymes, such as the gluconate dehydratase, the aldehyde oxidoreductase and GAPN, prevents the ED to operate in the gluconeogenic direction, which is, for instance, required to store energy in the form of glycogen (Skorko *et al.*, 1989). Another important role for the gluconeogenic EMP pathway is the production of fructose-6-phosphate (F6P), which has been proposed to be the main precursor for the Pentose Phosphate Pathway (PPP) (Verhees *et al.*, 2003). Except for three kinases (GK glucokinase, PFK phosphofructokinase and PK pyruvate kinase), the catabolic Embden-Meyerhof-Parnas (EMP) pathway consists of reversible enzymes. Although the genes encoding a GK and PFK were absent, the genes encoding the reversible EMP enzymes were all found in the genome of *Sulfolobus*. Moreover, a gene encoding a fructose-1,6-bisphosphatase (FBPase) was also detected. Because it is known that the catabolic EMP pathway is not operational in *Sulfolobus* (Selig *et al.*, 1997), it is likely that these EMP enzymes serve a gluconeogenic role. The simultaneous operation of both the ED and a gluconeogenic EMP pathway, however, requires a strict control of the metabolic flux through the pathway in order to prevent an energetically futile cycle. Allosteric regulation, post-translational protein modification and regulation at the transcriptional level are common strategies to modulate the activity and abundance of key enzymes, such as the fructose-1,6-bisphosphatase.

Although glycolysis in *Sulfolobus* is well studied, there are still unconfirmed genes and activities in the pathway. For instance, the transcriptome analysis revealed the expression of one of two putative gluconolactonases (Sso2705) that have generally been omitted in the analysis of the ED pathway, since the reaction from gluconolactone to gluconate also occurs spontaneously (Satory *et al.*, 1997). The expression of the enzyme, however, would suggest a functional role in the metabolism of *Sulfolobus*. Additionally, only one of two phosphoglycerate mutases (Sso0417) that were found in its genome was expressed in both the proteome and transcriptome, while the other type (Sso2236) remained undetected. Expression of the predicted glycerate kinase (Sso0666) was only detected at the mRNA level.

Tricarboxylic acid cycle

Sulfolobus spp. is an obligate aerobe that primarily obtains energy by the oxidation of organic molecules and elemental sulfur (Brock *et al.*, 1972). This oxidation results in the formation of reduced electron carriers, such as NAD(P)H, Fd_R and $FADH_2$. The majority of these reducing equivalents are generated in the tricarboxylic acid (TCA) cycle. Per round of the cycle, the succinate-CoA ligase of *Sulfolobus* generates one molecule of ATP, instead of the commonly produced GTP (Danson *et al.*, 1985). Apart from being the main metabolic converter of chemical energy, the TCA cycle intermediates serve an important role as biosynthetic precursors for many cellular components, such as amino acids. Consequently, when too many intermediates are withdrawn from the cycle, they need to be replenished by anaplerotic enzyme reactions. The phospho*enol*pyruvate carboxylase (PEPC), which forms oxaloacetate from phospho*enol*pyruvate, is the only anaplerotic enzyme from *Sulfolobus* that has been described to date (Ettema *et al.*, 2004; Sako *et al.*, 1996). A gene product with high similarity to known pyruvate carboxylases could not be detected in the predicted proteome of *Sulfolobus*. In the glyoxylate shunt, which is normally only active during growth on acetate, isocitrate and acetyl-CoA are converted into succinate and malate by the action of the isocitrate lyase and the malate synthase. Interestingly, the isocitrate lyase of glucose-grown *S. acidocaldarius* cells co-purified with the aconitase (Uhrigshardt *et al.*, 2001; Uhrigshardt *et al.*, 2002). Not only would this suggest a cytosolic association of the enzymes, but it also suggests that the glyoxylate shunt operates under saccharolytic conditions. This pathway may therefore constitute another way of replenishing four-carbon TCA cycle intermediates.

When there is an excess of TCA intermediates, for instance during growth on proteinaceous substrates, both malate and oxaloacetate can be decarboxylated to pyruvate by the malic enzyme (Bartolucci *et al.*, 1987). Oxaloacetate can also be converted to phosphoenolpyruvate by the GTP-dependent carboxykinase (Fukuda *et al.*, 2004). These four-to-three carbon conversions then provide the precursors that are required in, for instance, the gluconeogenesis pathway. In contrast to aerobic bacteria and eukaryotes, *Sulfolobus* uses ferredoxin instead of NAD$^+$ as a cofactor in the formation of acetyl-CoA from pyruvate and succinyl-CoA from 2-oxoglutarate (Kerscher *et al.*, 1982). The protein complex responsible for both conversions was shown to consist of two subunits; a fused α/γ subunit (Sso2815) and a β subunit (Sso2816) (Fukuda and Wakagi, 2002; Zhang *et al.*, 1996). The genome sequences of the three *Sulfolobus* species, however, revealed several paralogs of ferredoxin-dependent 2-oxoacid oxidoreductases, which might also be involved in these conversions.

What is also evident from this reconstruction is that almost all dehydrogenases in the central carbon metabolism of *Sulfolobus* show a clear cofactor preference for NADP$^+$ over NAD$^+$ (Bartolucci *et al.*, 1987; Camacho *et al.*, 1995; Danson *et al.*, 1985; Lamble *et al.*, 2003; Russo *et al.*, 1995; She *et al.*, 2001). The only exception to this rule seems to be the malate dehydrogenase, which, at least *in vitro*, uses both electron acceptors equally well (Hartl *et al.*, 1987). In bacteria and eukaryotes, most NADPH is usually formed in the PPP and used for reductive biosynthesis purposes. In *Sulfolobus*, the apparent enzyme preference for NADP$^+$ would suggest a more general role of its reduced form, in energy conservation by oxidative phosphorylation. Interestingly, as noted by She *et al.* (She *et al.*, 2001), all genes encoding the NAD(P)H dehydrogenase complex are present in the genome, except the three that encode the subunits which are required for NAD(P)H binding and oxidation. It has been proposed that the reducing equivalents are first transferred to ferredoxin by a NADPH:ferredoxin oxidoreductase, before entering the respiratory chain (She *et al.*, 2001).

Regulation of the main metabolic pathways

Insight was obtained into the regulation of the genes anticipated in glycolysis, gluconeogenesis and TCA cycle by measuring the relative abundance of their mRNA and protein levels by using a whole-genome DNA microarray and a quantitative proteomics approach, respectively (Table 2.1). In the measurements, 35 out of 41 transcripts ratios were determined, while 29 out of 41 protein ratios were analysed on 2-DE gels. On average the proteomic and transcriptomic data correlate reasonably well. For 26 genes both proteomic data and transcriptomic data are presented. In general, changes at proteomic and transcriptomic level show a similar trend, however, proteomic changes tend to be more pronounced. In only 3 cases the proteomic data contradict the transcriptome data. This concerns the three subunits for aldehyde dehydrogenase (Sso2639, Sso2636 and Sso2637). However, the fact that these clustered genes show a similar ratio at proteomic or transcriptomic level indicates the reliability of the data. Interestingly, all three subunits were found in the same protein spot on the gel, suggesting that a strong (non-covalent) interaction exists between them. The stability of the protein complex might be affected by stabilizing factors such as cofactors that may lead to different degrees of aggregation under different growth conditions. In terms of regulatory effects, the glyceraldehyde-3-phosphate dehydrogenase (non-phosphorylating; GAPN) was up-regulated under glucose conditions, or alternatively, down-regulated during growth in YT media. This is not surprising, since GAP is the crucial intermediate between the ED and gluconeogenic EMP, and too much of the strictly catabolic GAPN would be likely to interfere with gluconeogenesis. The enzymes involved in gluconeogenesis were all slightly up-regulated during growth on YT media, in agreement with expectations. Especially the phospho*enol*pyruvate synthase and the phosphoglycerate kinase, key enzymes of the pathway appeared to be most differentially expressed.

The expression levels of the TCA-cycle genes were only marginally different under the two conditions. Under glucose conditions, several enzymes of the TCA cycle were slightly induced at proteomic level, including the 2-oxoacid:ferredoxin oxidoreductase, the succinate-CoA ligase, the succinate dehydrogenase and the malate dehydrogenase. This was also true for the enzymes that replenish the four-carbon TCA cycle intermediates, such as the isocitrate lyase and the phospho*enol*pyruvate carboxylase. This ensures that sufficient oxaloacetate is present to serve as biosynthetic precursor and as an acceptor molecule for acetyl-CoA. The differences may be due to the fact that glucose catabolism mainly results in acetyl-CoA and oxaloacetate formation, whereas peptide degradation probably yields various central intermediates of carbon metabolism, such as pyruvate (Ala, Cys, Trp, Thr, Ser, Gly), acetyl-CoA (Phe, Tyr, Ile, Leu, Lys, Trp, Thr), 2-oxoglutarate (Arg, Gln, His, Pro, Glu), succinyl-CoA (Ile, Met, Val, Thr), fumarate (Phe, Tyr, Asp) and oxaloacetate (Asn, Asp).

Concluding remarks

In this study, we have created a proteome reference map for *Sulfolobus solfataricus* consisting of 325 proteins in 255 spots, and have reconstructed its central carbon metabolic pathways. The expression of the genes in these pathways was analysed by measuring the relative abundance of mRNA and protein under peptide- or sugar-degrading conditions. Although most observed differences were small, the expression of some key enzymes in glycolysis, gluconeogenesis and TCA cycle was significantly altered. Apart from looking at abundance levels, proteomics studies are now ongoing that focus on the modulation of enzyme activity by protein post-translational modification. These studies will provide additional clues that will reveal the details of regulation of the central carbon metabolism in *Sulfolobus solfataricus*.

Acknowledgements

The authors wish to thank: Anders Andersson and Peter Nilsson (Department of Biotechnology, Royal Institute of Technology Stockholm, Sweden) Magnus Lundgren and Rolf Bernander (Department of Molecular Evolution, Evolutionary Biology Center, Uppsala University, Uppsala, Sweden) for supplying the micro-array, technical assistance and advice. We thank the University of Sheffield and UK's SRIF Infrastructure funds for support. APLS thanks the University of Sheffield and the EPSRC for a scholarship. MGJV thanks the European Union's ERASMUS Program. PCW thanks the UK's Engineering and Physical Sciences Research Council (EPSRC) for provision of an Advanced Research Fellowship (GR/A11311/01). This work was partly supported by a grant from the European Union in the framework of the SCREEN project (contract QLK3-CT-2000-00649).

Supplementary material

For supplementary material see:

http://www.wiley-vch.de/contents/jc_2120/2006/pro2070_s.pdf.

References

Adams, M. W., Holden, J. F., Menon, A. L., Schut, G. J., Grunden, A. M., Hou, C., Hutchins, A. M., Jenney, F. E., Jr., Kim, C., Ma, K., *et al.* (2001). Key role for sulfur in peptide metabolism and in regulation of three hydrogenases in the hyperthermophilic archaeon Pyrococcus furiosus. J Bacteriol *183*, 716-724.

Ahmed, H., Ettema, T. J., Tjaden, B., Geerling, A. C., van der Oost, J., and Siebers, B. (2005). The semi-phosphorylative Entner-Doudoroff pathway in hyperthermophilic archaea: a re-evaluation. Biochem J *390*, 529-540.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997).

Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Andersson, A., Bernander, R., and Nilsson, P. (2005). Dual-genome primer design for construction of DNA microarrays. Bioinformatics *21*, 325-332.

Bartolucci, S., Rella, R., Guagliardi, A., Raia, C. A., Gambacorta, A., De Rosa, M., and Rossi, M. (1987). Malic enzyme from archaebacterium Sulfolobus solfataricus. Purification, structure, and kinetic properties. J Biol Chem *262*, 7725-7731.

Brinkman, A. B., Bell, S. D., Lebbink, R. J., de Vos, W. M., and van der Oost, J. (2002). The Sulfolobus solfataricus Lrp-like protein LysM regulates lysine biosynthesis in response to lysine availability. J Biol Chem *277*, 29537-29549.

Brock, T. D., Brock, K. M., Belly, R. T., and Weiss, R. L. (1972). Sulfolobus: a new genus of sulfur-oxidizing bacteria living at low pH and high temperature. Arch Mikrobiol *84*, 54-68.

Brunner, N. A., Brinkmann, H., Siebers, B., and Hensel, R. (1998). NAD+-dependent glyceraldehyde-3-phosphate dehydrogenase from Thermoproteus tenax. The first identified archaeal member of the aldehyde dehydrogenase superfamily is a glycolytic enzyme with unusual regulatory properties. J Biol Chem *273*, 6149-6156.

Buchanan, C. L., Connaris, H., Danson, M. J., Reeve, C. D., and Hough, D. W. (1999). An extremely thermostable aldolase from Sulfolobus solfataricus with specificity for non-phosphorylated substrates. Biochem J *343 Pt 3*, 563-570.

Camacho, M. L., Brown, R. A., Bonete, M. J., Danson, M. J., and Hough, D. W. (1995). Isocitrate dehydrogenases from Haloferax volcanii and Sulfolobus solfataricus: enzyme purification, characterisation and N-terminal sequence. FEMS Microbiol Lett *134*, 85-90.

Chen, L., Brugger, K., Skovgaard, M., Redder, P., She, Q., Torarinsson, E., Greve, B., Awayez, M., Zibat, A., Klenk, H. P., and Garrett, R. A. (2005). The genome of Sulfolobus acidocaldarius, a model organism of the Crenarchaeota. J Bacteriol *187*, 4992-4999.

Colombo, S., Grisa, M., Tortora, P., and Vanoni, M. (1994). Molecular cloning, nucleotide sequence and expression of a Sulfolobus solfataricus gene encoding a class II fumarase. FEBS Lett *337*, 93-98.

Contursi, P., Cannio, R., Prato, S., Fiorentino, G., Rossi, M., and Bartolucci, S. (2003). Development of a genetic system for hyperthermophilic Archaea: expression of a moderate thermophilic bacterial alcohol dehydrogenase gene in Sulfolobus solfataricus. FEMS Microbiol Lett *218*, 115-120.

Danson, M. J. (1988). Archaebacteria: the comparative enzymology of their central metabolic pathways. Adv Microb Physiol *29*, 165-231.

Danson, M. J., Black, S. C., Woodland, D. L., and Wood, P. A. (1985). Citric acid cycle enzymes of the archaebacteria: citrate synthase and succinate thiokinase. FEBS *179*, 120-124.

De Rosa, M., Gambacorta, A., Nicolaus, B., Giardina, P., Poerio, E., and Buonocore, V. (1984). Glucose metabolism in the extreme thermoacidophilic archaebacterium Sulfolobus solfataricus. Biochem J *224*, 407-414.

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

Ettema, T. J., Makarova, K. S., Jellema, G. L., Gierman, H. J., Koonin, E. V., Huynen, M. A., de Vos, W. M., and van der Oost, J. (2004). Identification and functional verification of archaeal-type phosphoenolpyruvate carboxylase, a missing link in archaeal central carbohydrate metabolism. J Bacteriol *186*, 7754-7762.

Fukuda, E., and Wakagi, T. (2002). Substrate recognition by 2-oxoacid:ferredoxin oxidoreductase from Sulfolobus sp. strain 7. Biochim Biophys Acta *1597*, 74-80.

Fukuda, W., Fukui, T., Atomi, H., and Imanaka, T. (2004). First characterization of an archaeal GTP-dependent phosphoenolpyruvate carboxykinase from the hyperthermophilic archaeon Thermococcus kodakaraensis KOD1. J Bacteriol *186*, 4620-4627.

Giometti, C. S., Reich, C., Tollaksen, S., Babnigg, G., Lim, H., Zhu, W., Yates, J., and Olsen, G. (2002). Global analysis of a “simple” proteome: Methanococcus jannaschii. J Chromatogr B Analyt Technol Biomed Life Sci *782*, 227-243.

Hansen, T., Wendorff, D., and Schonheit, P. (2004). Bifunctional phosphoglucose/phosphomannose isomerases from the Archaea Aeropyrum pernix and Thermoplasma acidophilum constitute a novel enzyme family within the phosphoglucose isomerase superfamily. J Biol Chem *279*, 2262-2272.

Hartl, T., Grossebuter, W., Gorisch, H., and Stezowski, J. J. (1987). Crystalline NAD/NADP-dependent malate dehydrogenase; the enzyme from the thermoacidophilic archaebacterium Sulfolobus acidocaldarius. Biol Chem Hoppe Seyler *368*, 259-267.

Hess, D., Kruger, K., Knappik, A., Palm, P., and Hensel, R. (1995). Dimeric 3-phosphoglycerate kinases from hyperthermophilic Archaea. Cloning, sequencing and expression of the 3-phosphoglycerate kinase gene of Pyrococcus woesei in Escherichia coli and characterization of the protein. Structural and functional comparison with the 3-phosphoglycerate kinase of Methanothermus fervidus. Eur J Biochem *233*, 227-237.

Hutchins, A. M., Holden, J. F., and Adams, M. W. (2001). Phosphoenolpyruvate synthetase from the hyperthermophilic archaeon Pyrococcus furiosus. J Bacteriol *183*, 709-715.

Huynen, M. A., Dandekar, T., and Bork, P. (1999). Variation and evolution of the citric-acid cycle: a genomic perspective. Trends Microbiol *7*, 281-291.

Janssen, S., Schafer, G., Anemuller, S., and Moll, R. (1997). A succinate dehydrogenase with novel structure and properties from the hyperthermophilic archaeon Sulfolobus acidocaldarius: genetic and biophysical characterization. J Bacteriol *179*, 5560-5569.

Jonuscheit, M., Martusewitsch, E., Stedman, K. M., and Schleper, C. (2003). A reporter gene system for the hyperthermophilic archaeon Sulfolobus solfataricus based on a selectable and integrative shuttle vector. Mol Microbiol *48*, 1241-1252.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res *32*, D277-280.

Kardinahl, S., Schmidt, C. L., Hansen, T., Anemuller, S., Petersen, A., and Schafer, G. (1999). The strict molybdate-dependence of glucose-degradation by the thermoacidophile Sulfolobus acidocaldarius reveals the first crenarchaeotic molybdenum containing enzyme--an aldehyde oxidoreductase. Eur J Biochem *260*, 540-548.

Kawarabayasi, Y., Hino, Y., Horikawa, H., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., *et al.* (2001). Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, Sulfolobus tokodaii strain7. DNA Res *8*, 123-140.

Kerscher, L., Nowitzki, S., and Oesterhelt, D. (1982). Thermoacidophilic archaebacteria contain bacterial-type ferredoxins acting as electron acceptors of 2-oxoacid:ferredoxin oxidoreductases. Eur J Biochem *128*, 223-230.

Kim, S., and Lee, S. B. (2005). Identification and characterization of Sulfolobus solfataricus D-gluconate dehydratase: a key enzyme in the non-phosphorylated Entner-Doudoroff pathway. Biochem J *387*, 271-280.

Kohlhoff, M., Dahm, A., and Hensel, R. (1996). Tetrameric triosephosphate isomerase from hyperthermophilic Archaea. FEBS Lett *383*, 245-250.

Lamble, H. J., Heyer, N. I., Bull, S. D., Hough, D. W., and Danson, M. J. (2003). Metabolic pathway promiscuity in the archaeon Sulfolobus solfataricus revealed by studies on glucose dehydrogenase and 2-keto-3-deoxygluconate aldolase. J Biol Chem *278*, 34066-34072.

Lamble, H. J., Milburn, C. C., Taylor, G. L., Hough, D. W., and Danson, M. J. (2004). Gluconate dehydratase from the promiscuous Entner-Doudoroff pathway in Sulfolobus solfataricus. FEBS Lett *576*, 133-136.

Limauro, D., Cannio, R., Fiorentino, G., Rossi, M., and Bartolucci, S. (2001). Identification and molecular characterization of an endoglucanase gene, celS, from the extremely thermophilic archaeon Sulfolobus solfataricus. Extremophiles *5*, 213-219.

Lohlein-Werhahn, G., Goepfert, P., and Eggerer, H. (1988). Purification and properties of an archaebacterial enzyme: citrate synthase from Sulfolobus solfataricus. Biol Chem Hoppe Seyler *369*, 109-113.

Lundgren, M., Andersson, A., Chen, L., Nilsson, P., and Bernander, R. (2004). Three replication origins in Sulfolobus species: synchronous initiation of chromosome replication and asynchronous termination. Proc Natl Acad Sci U S A *101*, 7046-7051.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., *et al.* (2005). InterPro, progress and status in 2005. Nucleic Acids Res *33*, D201-205.

Nishimasu, H., Fushinobu, S., Shoun, H., and Wakagi, T. (2004). The first crystal structure of the novel class of fructose-1,6-bisphosphatase present in thermophilic archaea. Structure (Camb) *12*, 949-959.

Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov, E., Jr., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I., *et al.* (2003). The ERGO genome analysis and discovery system. Nucleic Acids Res *31*, 164-171.

Pappin, D. J., Hojrup, P., and Bleasby, A. J. (1993). Rapid identification of proteins by peptide-mass fingerprinting. Curr Biol *3*, 327-332.

Peak, M. J., Peak, J. G., Stevens, F. J., Blamey, J., Mai, X., Zhou, Z. H., and Adams, M. W. (1994). The hyperthermophilic glycolytic enzyme enolase in the archaeon, Pyrococcus furiosus: comparison with mesophilic enolases. Arch Biochem Biophys *313*, 280-286.

Peterson, J. D., Umayam, L. A., Dickinson, T., Hickey, E. K., and White, O. (2001). The Comprehensive Microbial Resource. Nucleic Acids Res *29*, 123-125.

Puchegger, S., Redl, B., and Stoffler, G. (1990). Purification and properties of a thermostable fumarate hydratase from the archaeobacterium Sulfolobus solfataricus. J Gen Microbiol *136*, 1537-1541.

Ronimus, R. S., and Morgan, H. W. (2003). Distribution and phylogenies of enzymes of the Embden-Meyerhof-Parnas pathway from archaea and hyperthermophilic bacteria support a gluconeogenic origin of metabolism. Archaea *1*, 199-221.

Russo, A. D., Rullo, R., Masullo, M., Ianniciello, G., Arcari, P., and Bocchini, V. (1995). Glyceraldehyde-3-phosphate dehydrogenase in the hyperthermophilic archaeon Sulfolobus solfataricus: characterization and significance in glucose metabolism. Biochem Mol Biol Int *36*, 123-135.

Sako, Y., Takai, K., Uchida, A., and Ishida, Y. (1996). Purification and characterization of phosphoenolpyruvate carboxylase from the hyperthermophilic archaeon Methanothermus sociabilis. FEBS Lett *392*, 148-152.

Satory, M., Furlinger, M., Haltrich, D., Kulbe, K. D., Pittner, F., and Nidetzky, B. (1997). Continuous enzymatic production of lactobionic acid using glucose-fructose oxidoreductase in an ultrafiltration membrane bioreactor. Biotech Lett *19*, 1205-1208.

Schafer, G. (1996). Bioenergetics of the archaebacterium Sulfolobus. Biochim Biophys Acta *1277*, 163-200.

Schomburg, I., Chang, A., Ebeling, C., Gremse, M., Heldt, C., Huhn, G., and Schomburg, D. (2004). BRENDA, the enzyme database: updates and major new developments. Nucleic Acids Res *32*, D431-433.

Schonheit, P., and Schafer, G. (1995). Metabolism of hyperthermophiles. World J Microbiol Biotechnol *11*, 26-54.

Schramm, A., Siebers, B., Tjaden, B., Brinkmann, H., and Hensel, R. (2000). Pyruvate kinase of the hyperthermophilic crenarchaeote Thermoproteus tenax: physiological role and phylogenetic aspects. J Bacteriol *182*, 2001-2009.

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

Selig, M., Xavier, K. B., Santos, H., and Schonheit, P. (1997). Comparative analysis of Embden-Meyerhof and Entner-Doudoroff glycolytic pathways in hyperthermophilic archaea and the bacterium Thermotoga. Arch Microbiol *167*, 217-232.

She, Q., Singh, R. K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M. J., Chan-Weiher, C. C., Clausen, I. G., Curtis, B. A., De Moors, A*., et al.* (2001). The complete genome of the crenarchaeon Sulfolobus solfataricus P2. Proc Natl Acad Sci U S A *98*, 7835-7840.

Siebers, B., Brinkmann, H., Dorr, C., Tjaden, B., Lilie, H., van der Oost, J., and Verhees, C. H. (2001). Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase. J Biol Chem *276*, 28710-28718.

Skorko, R., Osipiuk, J., and Stetter, K. O. (1989). Glycogen-bound polyphosphate kinase from the archaebacterium Sulfolobus acidocaldarius. J Bacteriol *171*, 5162-5164.

Smith, L. D., Stevenson, K. J., Hough, D. W., and Danson, M. J. (1987). Citrate synthase from the thermophilic archaebacteria Thermoplasma acidophilum and Sulfolobus acidocaldarius. FEBS Lett *225*, 277-281.

Snijders, A. P., de Vos, M. G., de Koning, B., and Wright, P. C. (2005a). A fast method for quantitative proteomics based on a combination between two-dimensional electrophoresis and (15)N-metabolic labelling. Electrophoresis *26*, 3191-3199.

Snijders, A. P., de Vos, M. G., and Wright, P. C. (2005b). Novel approach for peptide quantitation and sequencing based on 15N and 13C metabolic labeling. J Proteome Res *4*, 578-585.

Solow, B., Bischoff, K. M., Zylka, M. J., and Kennelly, P. J. (1998). Archael phosphoproteins. Identification of a hexosephosphate mutase and the α-subunit of succinyl-CoA synthetase in the extreme acidothermophile Sulfolobus solfataricus. Protein Sci *7*, 105-111.

Stedman, K. M., Schleper, C., Rumpf, E., and Zillig, W. (1999). Genetic requirements for the function of the archaeal virus SSV1 in Sulfolobus solfataricus: construction and testing of viral shuttle vectors. Genetics *152*, 1397-1405.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N*., et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

Tusher, V. G., Tibshirani, R., and Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. Proc Natl Acad Sci U S A *98*, 5116-5121.

Uhrigshardt, H., Walden, M., John, H., and Anemuller, S. (2001). Purification and characterization of the first archaeal aconitase from the thermoacidophilic Sulfolobus acidocaldarius. Eur J Biochem *268*, 1760-

1771.

Uhrigshardt, H., Walden, M., John, H., Petersen, A., and Anemuller, S. (2002). Evidence for an operative glyoxylate cycle in the thermoacidophilic crenarchaeon Sulfolobus acidocaldarius. FEBS Lett *513*, 223-229.

Van der Oost, J., Huynen, M. A., and Verhees, C. H. (2002). Molecular characterization of phosphoglycerate mutase in archaea. FEMS Microbiol Lett *212*, 111-120.

Verhees, C. H., Kengen, S. W., Tuininga, J. E., Schut, G. J., Adams, M. W., De Vos, W. M., and Van Der Oost, J. (2003). The unique features of glycolytic pathways in Archaea. Biochem J *375*, 231-246.

Worthington, P., Hoang, V., Perez-Pomares, F., and Blum, P. (2003). Targeted Disruption of the α-Amylase Gene in the Hyperthermophilic Archaeon Sulfolobus solfataricus. J Bacteriol *185*, 482-488.

Zaigler, A., Schuster, S. C., and Soppa, J. (2003). Construction and usage of a onefold-coverage shotgun DNA microarray to characterize the metabolism of the archaeon Haloferax volcanii. Mol Microbiol *48*, 1089-1105.

Zhang, Q., Iwasaki, T., Wakagi, T., and Oshima, T. (1996). 2-oxoacid:ferredoxin oxidoreductase from the thermoacidophilic archaeon, Sulfolobus sp. strain 7. J Biochem (Tokyo) *120*, 587-599.

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

Chapter 3

Pentose metabolism in archaea

van de Werken, H. J. G., Brouns, S. J. J., and van der Oost, J. (2008). In Archaea, P. Blum, ed. (Wymondham, Caister Academic Press).

Abstract

Archaeal physiology has been studied extensively ever since the discovery that they constitute a distinct domain of life. The diversity of the archaeal metabolism is very high: they are able to grow fermentatively, but are also able to respire aerobically or anaerobically, and obtain their energy from light or (an)organic molecules. Initially, hexose conversions received most attention. Recently, however, significant insight has been gained in the archaeal pentose metabolism. Importantly, novel genomic, genetic and bioinformatic tools are applicable now to study the archaeal biology in more detail. We here compare the archaeal pentose pathways, the enzymes and their regulation to the bacterial and eukaryal counterparts, and also describe distinct archaeal metabolic features and their implications. The pentose metabolism in Archaea shows a mosaic of both bacterial and eukaryal anabolic and catabolic pathways, but has also unique archaeal conversions, novel enzymes and distinct regulatory features. This reflects the archaeal evolution of a variable metabolic shell that adjusts to the availability of the substrates and the extreme conditions under which many Archaea live.

Introduction

Since their discovery and classification Archaea have been recognized as a third domain of life, phylogenetically distinct from Bacteria and Eukarya (Woese and Fox, 1977; Woese *et al.*, 1990). Eukaryal organisms are clearly different from Bacteria and Archaea, because of their more complex structural organization: without exception they possess intracellular compartments and in addition they often have a multicellular composition. The morphology of the prokaryotic Archaea and Bacteria is very similar, and by definition, these unicellular organisms do not have a nucleus, or have any other cytoplasmic compartments. Because of this similarity, the co-existence of two fundamentally different types of prokaryotes was not recognized before the introduction of molecular classification techniques in the 1970s. These molecular analyses strongly suggested that early in the cellular evolution two domains diverged within the prokaryotes: the Archaea (Archaebacteria) and the Bacteria (Eubacteria). The comparison of complete genomes that were released during the last decade, generally has confirmed the proposed division into three monophylic domains (Ciccarelli *et al.*, 2006; Snel *et al.*, 1999).

Although the majority of the Archaea was initially isolated from extreme environments (high temperature, high salt concentration, extreme pH), it has become clear that Archaea also thrive in non-extreme environments. Moreover, members of this domain of life are abundant in many ecosystems ranging from soil to marine environments (Pace, 1997; Schleper *et al.*, 2005; Sinninghe Damste *et al.*, 2002). Representatives of the archaeal classes discovered several decades ago (methanogen Methanocaldococcus jannaschii, halophile Halobacterium salinarum and (hyper)thermophiles, such as Pyrococcus furiosus and Sulfolobus solfataricus) have become model organisms for studying archaeal metabolism (physiology, biochemistry, genetics and genomics).

The metabolism of the Bacteria and Eukarya has been studied in great detail; especially the composition and capacity of the bacterial metabolic system is very versatile. Although the metabolism of Archaea has been investigated to a much lesser extent, it is clear that also their metabolism is very diverse. It ranges from fermentation to anaerobic and aerobic respiration and from photo- and chemolithotrophy to heterotrophy (Schonheit and Schafer, 1995). The archaeal hexose metabolism has been extensively studied in the last decades. Archaea degrade the hexose glucose via the modified Embden–Meyerhof–Parnas (EMP) or Entner–Doudoroff (ED) pathways (De Rosa *et al.*, 1984) and use distinct control mechanisms (for reviews see (Siebers and Schonheit, 2005; van der Oost and Siebers, 2007; Verhees *et al.*, 2003)). Pentose metabolism, however, has only recently been addressed in Archaea (Brouns *et al.*, 2006; Johnsen and Schonheit, 2004).

The aim of this chapter is to summarize some recent discoveries on the pentose metabolism of Archaea. Relevant pentose-converting routes, including novel enzymes and unique regulatory mechanisms, are compared to Bacteria and Eukarya.

Archaeal pentose anabolic metabolism

Pentose phosphate pathway

The pentose phosphate pathway (PPP) in Bacteria and Eukarya has a dual function. Firstly, it generates reducing power (NADPH) that serves as an electron donor in biosynthesis pathways. Secondly, it provides ribose-5-phosphate as building blocks for nucleotides and as precursor for histidine and coenzyme biosynthesis (riboflavin and NAD$^+$) and erythrose-4-phosphate for the synthesis of the aromatic amino acids (Fig. 3.1).

The PPP can be divided in an oxidative (OPPP) and a non-oxidative branch (NOPPP). The OPPP involves the step-wise oxidation of glucose-6-phosphate to ribulose-5-phosphate accompanied by the formation of two NADPH. The NOPPP links the glycolysis with the PPP by the transketolase and the transaldolase (Fig. 3.1). Based on comparative genomics, it has been proposed that ribose-5-phosphate can be converted to fructose-6-phosphate and glyceraldehyde-3-phosphate, by both enzymes, in Thermoplasmales, Methanococcales and *Cenarchaeum symbiosum* (for reviews see: (Soderberg, 2005; Verhees *et al.*, 2003)). Remarkably, no archaeal enzymes are known to catalyze the oxidative branch (see below), although candidates involved in 6-phosphogluconate oxidation have been proposed (Soderberg, 2005). In addition, it has been demonstrated that the methanogens *Methanococcus maripaludis*

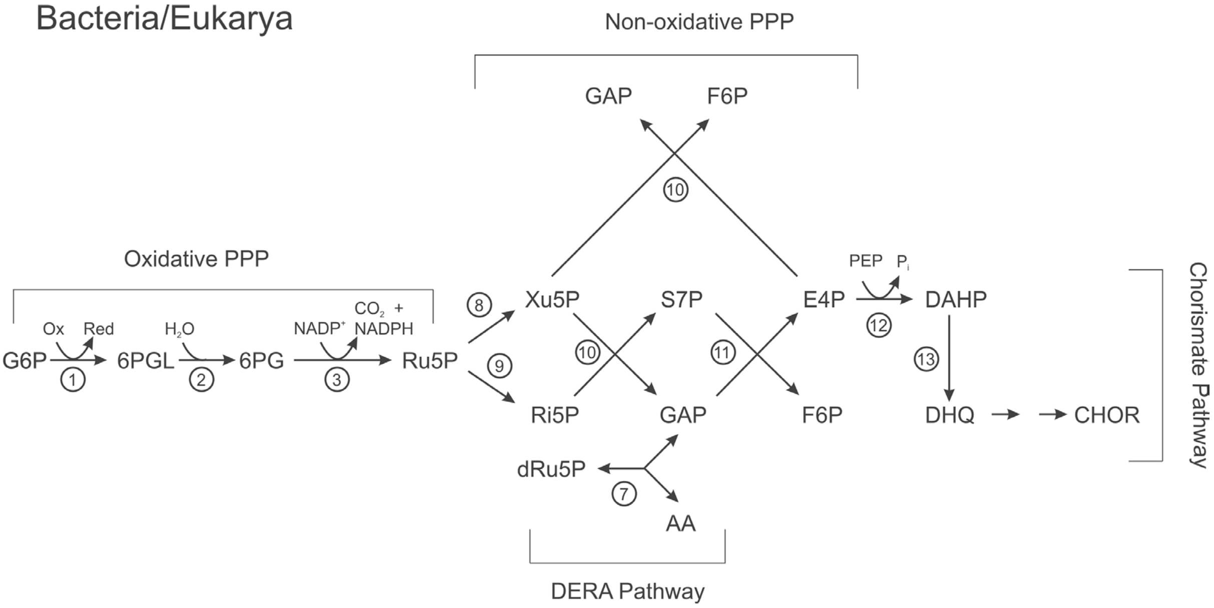


Figure 3.1 Pentose anabolic metabolism in Bacteria and Eukarya. Pathways that lead to the generation of pentoses and chorismate in Bacteria and Eukarya. Arrows represent enzymatic steps that are performed by known proteins described in Table 4.1. Abbreviated metabolites are: G6P, glucose-6-phosphate; 6PGL, 6-phosphoglucono-δ-lactone; 6PG, 6-phosphogluconate; Ru5P, ribulose-5-phosphate; GAP, glyceraldehyde-3-phosphate; AA, acetaldehyde; dRu5p, 2-deoxyribose-5-phosphate; Xu5p, xylulose-5-phosphate; Ri5P ribose-5-phosphate; S7P, sedoheptulose-7-phosphate; E4P, erythrose-4-phosphate; F6P; fructose-6-phosphate; PEP, phosphoenolpyruvate; Pi, inorganic phosphate; DAHP, 3-deoxy-d-arabino-2-heptulosonate-7-phosphate; DHQ, 3-dehydroquinate; CHOR chorismate. Pathway abbreviations are PPP, pentose phosphate pathway; DERA, 2-deoxyribose-5-phosphate aldolase. The figure was adapted from (Orita *et al.*, 2006).

and *Methanococcus voltae* do not use the classical OPPP (Choquet *et al.*, 1994; Tumbula *et al.*, 1997; Yu *et al.*, 1994). Moreover, no OPPP enzyme activity has been measured in distinct methanogens, such as *Methanospirillum hungatei, Methanothermobacter thermautotrophicus* (formerly known as *Methanobacterium thermoautotrophicum*) Marburg and *Methanosarcina barkeri* Fusaro (Purwantini *et al.*, 1997) and in the crenarchaeon *Sulfolobus solfataricus* (De Rosa *et al.*, 1984).

A functional OPPP, however, can not be excluded for *M. hungatei, M. thermautotrophicus*, *Methanobrevibacter smithii, Methanosphaera stadtmanae*, *Methanosarcina barkeri* and *Methanobacterium bryantii*, since carbon labeling studies showed synthesis of labeled ribose-5-phosphate that was consistent with a functional OPPP (Choquet *et al.*, 1994; Eisenreich *et al.*, 1991; Ekiel *et al.*, 1983). Recently, White and co-workers showed that the PPP is not essential in *Methanocaldococcus jannaschii* (formerly known as *Methanococcus jannaschii*) for the production of ribose-5-phosphate and erythrose-4-phosphate (Grochowski *et al.*, 2005; White, 2004). This is in agreement with genome analysis that suggests either that the classical pathway is not present in methanogens, or that the enzymes involved are not orthologous to the bacterial and eukaryal counterparts.

The oxidative pentose phosphate pathway

Glucose-6-phosphate 1-dehydrogenase

This enzyme is responsible for the oxidation of glucose-6-phosphate to 6-phosphogluconolactone under the formation of one molecule of NADPH. This conversion is part of both the PPP and the ED pathway (Fig. 3.1/Table 3.1). NADP-dependent glucose-6-phosphate 1-dehydrogenase (G6PDH) activity has never been detected in cell-extracts, neither have orthologs of the bacterial G6PDH and eukaryal G6PDH (both belonging to the Clusters of Orthologous Groups of proteins COG0364 (Tatusov *et al.*, 2003)) been found in archaeal genomes sequenced to date.

Also the F_{420}-dependent glucose-6-phosphate dehydrogenase (FGD, COG2141) activity has not been determined in Archaea. However, close homologs of the *Mycobacterium smegmatis* FDG are identified in genome sequences in several methanogens and in *Archaeoglobus fulgidus* (Purwantini and Daniels, 1998). Unfortunately, the genomic neighborhood of these archaeal genes is not conserved strong enough to predict a reliable function for the F_{420}-dependent enzymes.

6-Phosphogluconolactonase

The 6-phosphogluconolactonase hydrolyzes the lactone ring, producing an aldonic acid. Currently, 6-phosphogluconolactonase (COG0363) has only been found in Bacteria and Eukarya. Archaeal. Enzymes that hydrolyze this lactone might be related to COG2220 or

COG3386 pentonolactonases (see: Clustering of genes involved in pentose oxidation). However, thermophilic Archaea may not require enzymatic lactonase activity, since these molecules are not very stable, and will hydrolyze spontaneously at non-limiting rates at high temperature (Brouns *et al.*, 2006).

6-Phosphogluconate dehydrogenase (decarboxylating)

The final step of the OPPP is the oxidation and decarboxylation of 6-phosphogluconate to ribulose-5-phosphate. At present, this step has only been demonstrated in Bacteria and Eukarya (COG0362, gnd), although it has been proposed that the genome of *Halobacterium salinarum* encodes a distantly related 6-phosphogluconate dehydrogenase (COG1023) (Soderberg, 2005). In addition, an even more distantly related orthologous group of proteins might be capable of catalyzing the same conversion within the Archaea (COG2084). Despite the fact that the genome sequences of Archaea lack a classical OPPP, it cannot be ruled out that non-homologous proteins are capable of catalyzing the similar conversion of 6-phosphogluconate. A second possibility is that some Archaea do possess a modified OPPP, which was not detected based on classical enzyme activity assays. In particular, Archaea that metabolize glucose via a modified ED pathway (*Sulfolobus* spp., *Thermoproteus* spp. and halophiles (Ahmed *et al.*, 2005)) might have the capacity to phosphorylate gluconate and subsequently catalyze the oxidative decarboxylation. Further research is needed to test these speculations.

The non-oxidative pentose phosphate pathway

The function of the non-oxidative pentose phosphate pathway is to convert pentoses to central metabolic intermediates such as glyceraldehyde-3-phosphate, fructose-6-phosphate, ribose-5-phosphate, and erythrose-4-phosphate, which serve as building blocks for the various biosynthesis pathways (Fig. 3.1). The interconversions of these compounds are accomplished by the classical enzymes ribose-5-phosphate isomerase, ribulose-phosphate 3-epimerase, transketolase and transaldolase (Table 3.1). Erythrose-4-phosphate is a precursor for 3-dehydroquinate (DHQ), which subsequently can be metabolized to chorismate (Fig. 3.1). DHQ can also be synthesized by a recently discovered 6-deoxy-5-ketofructose-1-phosphate pathway (see 6-deoxy-5-ketofructose-1-phosphate (DKFP) pathway, Fig. 3.2). The synthesis of DHQ in methanogens by the DKFP pathway is in agreement with ^{13}C-labeling experiments (Choquet *et al.*, 1994; Tumbula *et al.*, 1997; Yu *et al.*, 1994). Based on the genomic data, the NOPPP seems to be complete only in Thermoplasmales, Methanococcales and *Cenarchaeum symbiosum* (all species found in footnote of Table 3.1 were used for this analysis). Nevertheless, only two archaeal proteins, ribose-5-phosphate isomerase in *P. horikoshii* (Ishikawa *et al.*, 2002) and transaldolase in *Methanocaldococcus jannaschii* (Soderberg and Alver, 2004) have been characterized (Table 3.1).

Ribose-5-phosphate isomerase

The reversible isomerization reaction of ribulose-5-phosphate to ribose-5-phosphate can be catalyzed by two distinct protein families: the non-inducible RpiA (COG0120) and the inducible bacterial RpiB (COG0698). RpiB is absent in Archaea, but RpiA seems to be ubiquitous in all Archaea. This contrasts to the fact that only a small portion of the Archaea possesses the complete NOPPP. Thus, the ribose-5-phosphate isomerase is probably involved in the NOPPP, as well as in alternative pathways that lead to the production of riboses. The crystal structure of the tetrameric ribose-5-phosphate isomerase of *Pyrococcus horikoshii*, which is homologous to RpiA, has been solved (Ishikawa *et al.*, 2002).

Ribulose-phosphate 3-epimerase

This enzyme catalyzes the epimerization of ribulose-5-phosphate to xylulose-5-phosphate (X5P). Ribulose-phosphate 3-epimerase Rpe (COG0036) appears to be encoded only by the genomes of *Methanocaldococcus jannaschii, Methanococcus maripaludis, Thermoplasma acidophilum, Thermoplasma volcanium Picrophilus torridus* and *Cenarchaeum symbiosum*. In the latter case it is fused to the transketolase N-terminal subunit. Remarkably, these organisms encode also a transaldolase that is essential for the complete NOPPP.

Transketolase

To connect the pentoses R5P and X5P to the glycolysis the chemical compounds need to be converted to fructose-6-phosphate and glyceraldehyde-3-phosphate. These conversions need two transketolase and one transaldolase reactions that results in an overall formation of two fructose-6-phosphate molecules and one glyceraldehyde-3-phosphate out of three pentose molecules. The classical transketolase reactions in Bacteria and Eukarya are catalyzed by a single enzyme TktA (COG0021) (Table 3.1). In Archaea this enzyme appears to be the result of a gene-fission, creating an N- and C-terminal part (COG3958/3959). These enzymes are present in most Archaea that can synthesis aromatic amino acids, such as Sulfolobus solfataricus and Pyrococcus abyssi, while they are lacking in almost all organisms that possess the novel 6-deoxy-5-ketofructose-1-phosphate pathway (see below). This indicates that the former organisms use the classical chorismate pathway, which requires erythrose-4-phosphate and phosphoenolpyruvate as starting compounds. Interestingly, these genes are frequently clustered

Table 3.1 Pentose anabolic enzymes that are involved in pathways that lead to the generation of pentoses and 3-dehydroquinate in Archaea and their bacterial and eukaryal counterparts.

No	Protein ID	Protein Name	EC[a]	COG[b]	ABE[c]	Characterized in Archaea[d]	Reference
Oxidative Pentose Phosphate Pathway							
1	G6PDH/ Zwf	Glucose-6-phosphate 1-dehydrogenase	1.1.1.49	COG0364	BE		
1	FGD	F_{420}-dependent glucose-6-phosphate dehydrogenase	1.1.1.-	COG2141	AB		
2	6PGL	6-Phosphogluconolactonase	3.1.1.31	COG0363	BE		
3	Gnd	6-Phosphogluconate dehydrogenase(decarboxylating)	1.1.1.44	COG0362	BE		
3		Predicted 6-phosphogluconate dehydrogenase	1.1.1.44	COG1023	AB		
3		Predicted 6-phosphogluconate dehydrogenase	1.1.1.44	COG2084	AB		
Non-oxidative Pentose Phosphate Pathway							
8	Rpe	Ribulose-phosphate 3-epimerase	5.1.3.1	COG0036	ABE		
9	RpiA	Ribose-5-phosphate isomerase	5.3.1.6	COG0120	ABE	PH1375	(Ishikawa *et al.*, 2002)
9	RpiB	Ribose-5-phosphate isomerase	5.3.1.6	COG0698	B		
10	TktA	Transketolase	2.2.1.1	COG0021	BE		
10	Tkt1	N-terminal Transketolase	2.2.1.1	COG3958	AB		
10	Tkt2	C-terminal Transketolase	2.2.1.1	COG3959	AB		
11	TalB	Transaldolase	2.2.1.2	COG0176	ABE	MJ0960	(Soderberg and Alver, 2004)
Ribulose monophosphate pathway							
4	PGI	Glucose-6-phosphate isomerase	5.3.1.9	COG0166	ABE		
4	PGI	Glucose-6-phosphate isomerase	5.3.1.9	COG2140	A	PF0196	(Hansen *et al.*, 2001; Verhees *et al.*, 2001)
5	PHI	6-Phospho-3-hexuloisomerase	5.-.-.-	COG0794	AB	MJ1247	(Martinez-Cruz *et al.*, 2002)
6	HPS	3-Hexulose-6-phosphate synthase	4.1.2.-	COG0269	AB		
D-Deoxyribose 5-phosphate aldolase (DERA) pathway							
-	PPM	Phosphopentomutase	5.4.2.7	COG1109	ABE	TK1777	(Rashid *et al.*, 2004)
-	PPM	Phosphopentomutase	5.4.2.7	COG1015	B		
7	DERA	Deoxyribose-5-phosphate aldolase	4.1.2.4	COG0274	AB	TK2104/ APE2437	(Rashid *et al.*, 2004)/(Sakuraba *et al.,* 2003)
AMP-metabolism pathway							
-	DeoA	Thymidine phosphorylase/AMP phosphorylase	2.4.2.4/ 2.4.2.-	COG0213	AB	TK0352	(Sato *et al.*, 2007)
-	E2b2	Translation initiation factor 2B subunit/Ribose-1,5-bisphosphate isomerase	-/5.3.1.6	COG1184	ABE	TK0185	(Sato *et al.*, 2007)

-	RbcL	Ribulose-bisphosphate carboxylase type III	4.1.1.39	COG1850	AB	TK2290	(Sato *et al.*, 2007)
Canonical chorismate pathway for 3-dehydroquinate pathway							
12	AroA	3-Deoxy-7-phosphoheptulonate synthase	2.5.1.54	COG2876	ABE		
12	AroG	3-Deoxy-7-phosphoheptulonate synthase	2.5.1.54	COG3200	B		
12	AroG	3-Deoxy-7-phosphoheptulonate synthase	2.5.1.54	COG0722	BE		
13	AroB	3-Dehydroquinate synthase	4.2.3.4	COG0337	ABE		
6-Deoxy-5-ketofructose-1-phosphate (DKFP) pathway							
14	AroA’	2-Amino-3,7-dideoxy-D-threo-hept-6-ulosonic acid synthase	-	COG1830	AB	MJ0400/ MMP0686	(White, 2004)/ (Porat *et al.*, 2006)
15	AroB’	3-Dehydroquinate synthase II	-	COG1465	A	MJ1249/ MMP0006	(White, 2004)/ (Porat *et al.*, 2006)
-	FBA	DhnA-type fructose-1,6-bisphosphate aldolase and related enzymes/DKFP transaldolase	-	COG1830	AB	MJ1585	(Siebers *et al.*, 2001)

No represents conversion numbers in Figs 3.1 or 3.2.

[a]EC, Enzyme Commission (Barthelmes *et al.*, 2007); [b]COG, Clusters of Orthologous Groups of Proteins (Tatusov *et al.*, 2003); [c]Domains are A: Archaea, B: Bacteria and E: Eukarya. [d]Locus id. The archaeal genome analyses is based on complete genome sequences of Aeropyrum pernix, Hyperthermus butylicus, Sulfolobus acidocaldarius, Sulfolobus solfataricus, Sulfolobus tokodaii, Pyrobaculum aerophilum, Pyrobaculum islandicum, Thermofilum pendens, Archaeoglobus fulgidus, Haloarcula marismortui, Halobacterium sp. NRC-1, Haloquadratum walsbyi, Natronomonas pharaonis, Methanosphaera stadtmanae, Methanothermobacter thermautotrophicus, Methanocaldococcus jannaschii, Methanococcus maripaludis, Methanopyrus kandleri, Methanosaeta thermophila, Methanosarcina acetivorans, Methanosarcina barkeri, Methanosarcina mazei, Methanospirillum hungatei, Pyrococcus abyssi, Pyrococcus furiosus, Pyrococcus horikoshii, Thermococcus kodakarensis, Picrophilus torridus, Thermoplasma acidophilum, Thermoplasma volcanium and Nanoarchaeum equitans.

with amino acid biosynthesis genes, supporting their involvement in erythrose-4-phosphate synthesis (Verhees *et al.*, 2003).

Transaldolase

The transaldolase MJ0960 (COG0176) from *Methanocaldococcus jannaschii* has been experimentally characterized. It catalyzes the reaction of glyceraldehyde-3-phosphate and sedoheptulose-7-phosphate to form fructose-6-phosphate and erythrose-4-phosphate (Soderberg and Alver, 2004). The same reaction in Eukarya and Bacteria is involved in the recycling of the glycolytic compounds. In Archaea, the corresponding genes are only present in genomes that encode the complete NOPPP.

Concluding remarks on the non-oxidative PP pathway in Archaea

The overall picture of the NOPPP in Archaea is that only a few (Thermoplasmales, Methanococcales and *Cenarchaeum symbiosum*) are capable of completely interconverting glycolytic intermediates and pentoses. However, several Archaea, such as *Sulfolobus* spp. and *Pyrobaculum* spp. can generate DHQ for aromatic amino acid biosynthesis, but have an incomplete NOPPP and are probably using the ribulose monophosphate pathway (see: below). Apart from the ribose-5-phosphate isomerase, the other Archaea do not possess obvious homologs of the proteins that catalyze these conversions in Bacteria and Eukarya.

Ribulose monophosphate pathway (RuMP)

Some Archaea do use the reverse ribulose monophosphate pathway to metabolize ribose-5-phosphate (Fig. 3.2). This pathway is used in methylotrophic Bacteria to fix formaldehyde with ribulose-5-phosphate to d-arabino-3-hexulose-6-phosphate, which is subsequently isomerized to fructose-6-phosphate. In *Thermococcus kodakaraensis*, and most likely in more Archaea, the fused-enzyme 3-hexulose-6-phosphate synthase and 6-phospho-3-hexuloisomerase have been demonstrated to be involved in the reverse reaction, i.e. in the conversion of fructose-6-phosphate to ribulose-5-phosphate (Orita *et al.*, 2006). This conversion has previously been postulated to occur in *M. jannaschii* (Grochowski *et al.*, 2005). A comparative analysis shows that in those archaeal organisms that do not possess the complete NOPPP, the RuMP fills the gaps. Surprisingly, halophiles are lacking both pathways and may have evolved a different, yet unknown solution (Soderberg, 2005; Verhees *et al.*, 2003). For organisms that depend on RuMP to synthesize nucleotides, the final step is synthesized by ribose-5-phosphate isomerase (see: NOPPP), which is encoded in all completely sequenced archaeal genomes. Finally, this

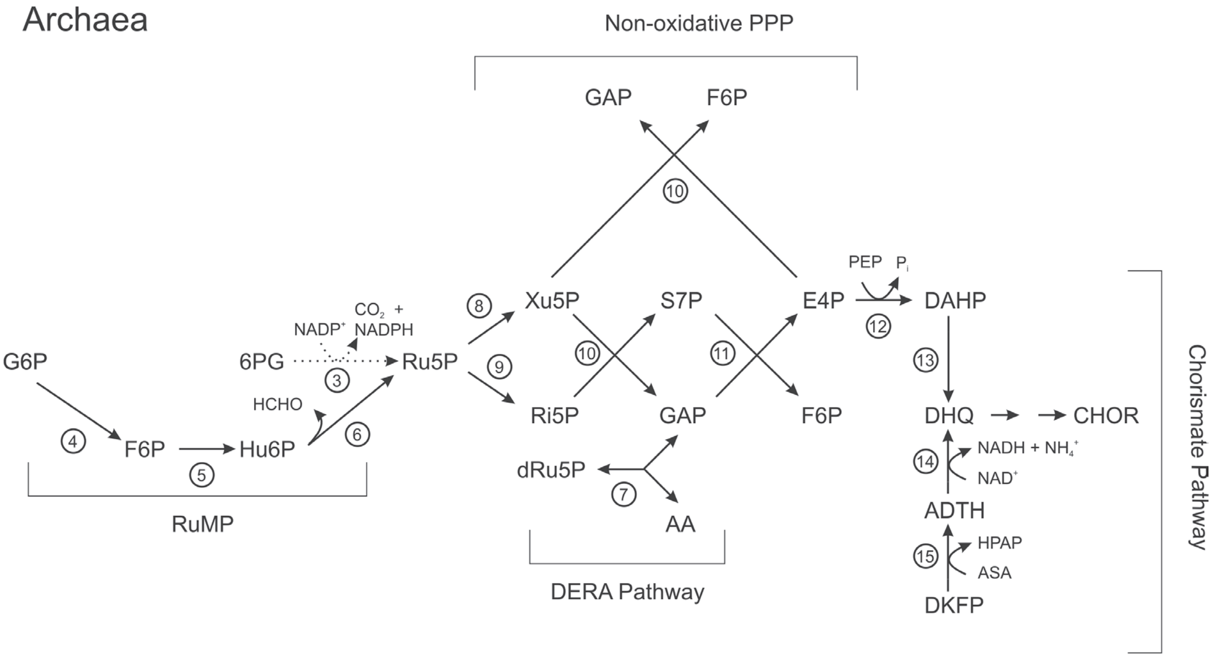


Figure 3.2 Pentose anabolic metabolism in Archaea. Pathways that lead to the generation of pentoses and chorismate in Archaea. Closed arrows represent enzymatic steps that are performed by known proteins, while dashed arrows are proteins with putative function; both are described in Table 4.1. Abbreviated metabolites are: G6P, glucose-6-phosphate; F6P, fructose-6-phosphate; Ru5P, ribulose-5-phosphate; GAP, glyceraldehyde-3-phosphate; AA, acetaldehyde; dRu5p, 2-deoxyribose-5-phosphate; Xu5p, xylulose-5-phosphate; Ri5P ribose-5-phosphate; S7P, sedoheptulose-7-phosphate; E4P, erythrose-4-phosphate; PEP, phosphoenolpyruvate; Pi, inorganic phosphate; DAHP, 3-deoxy-d-arabino-2-heptulosonate-7-phosphate; DHQ, 3-dehydroquinate; ADTH, 2-amino-3,7-dideoxy-d-threo-hept-6-ulosonic; DKFP, 6-deoxy-5-ketofructose-1-phosphate; ASA, l-aspartate semialdehyde; HPAP, hydroxypyruvaldehyde phosphate; CHOR chorismate. Pathway abbreviations are PPP, pentose phosphate pathway; RuMP ribulose monophosphate pathway; DERA, 2-deoxyribose-5-phosphate aldolase. The figure was adapted from Orita *et al.* (2006).

pathway is in good agreement with results of the ^{13}C-labeling experiments in *Thermococcus zilligii* (Xavier *et al.*, 2000).

6-Phospho-3-hexuloisomerase

The first step in synthesizing ribulose-5-phosphate is the isomerization of fructose-6-phosphate to D-*arabino*-3-hexulose-6-phosphate by 6-phospho-3-hexuloisomerase (PHI). In Thermococcales, the enzyme PHI (COG0794) is fused to 3-hexulose-6-phosphate synthase (HPS) and it originally was suggested to be involved in formaldehyde fixation in *P. horikoshii* (Orita *et al.*, 2005). However, genetic analyses revealed that this gene is essential for the synthesis of nucleosides in *Thermococcus kodakaraensis* (Orita *et al.*, 2006). The fusion may be beneficial for these hyperthermophiles, because D-*arabino*-3-hexulose-6-phosphate is unstable at elevated temperatures (Kemp, 1974). Recently the crystal structure of the *M. jannaschii* PHI (MJ1247) was determined (Martinez-Cruz *et al.*, 2002).

3-Hexulose-6-phosphate synthase

The second step, cleaving of formaldehyde from D-*arabino*-3-hexulose-6-phosphate, results in ribulose-5-phosphate. Besides being fused to PHI, HPS (COG0269)-containing organisms also possess an additional domain corresponding to either a formaldehyde-activating enzyme (Vorholt *et al.*, 2000) (in methanogens and in *Archaeoglobus*; COG1795), or a tungsten-dependent formaldehyde ferredoxin oxidoreductase (Roy *et al.*, 1999)) (in *Pyrococcus furiosus*, COG2414) to detoxify formaldehyde. In addition, Soderberg (2005) concludes that only the *Sulfolobus solfataricus* genome does not contain an open reading frame of the latter two COGs. However, SSO0472 (COG1062), a gluthatione-independent formaldehyde dehydrogenase is a good candidate to fill the gap. The enzyme is homologous to the characterized enzyme in *Pseudomonas putida* (Ogushi *et al.*, 1984), but might have a different function in *E. coli*, because it is fused to a Ribonuclease E (RNase E) inhibitor. Nevertheless, in many Archaea PHI and HPI are essential in generating pentoses and COG1062 is good candidate, in some Archaea, to remove the toxic formaldehyde.

Metabolic link between pentoses, chorismate and central carbon metabolism

2-Deoxyribose 5-phosphate aldolase (DERA) pathway

In many Bacteria and Eukarya the link between the central carbohydrate metabolism and nucleosides is the conversion from acetaldehyde and glyceraldehyde-3-phosphate to pentoses. This reversible pathway is catalyzed by a 2-deoxyribose-5-phosphate aldolase (DERA, Fig. 3.1) and a phosphopentomutase (PPM) to produce (deoxy)ribose 1-phosphate, which can be used as ribose moiety of nucleosides by nucleoside phosphorylases. A close homolog of the bacterial 2-deoxyribose-5-phosphate aldolase (DeoC/DERA; TK2104/APE2437, COG0274) was identified and characterized in *T. kodakaraensis* (Rashid *et al.*, 2004) and the crystal structure of the ortholog from *Archaeoglobus fulgidus* was solved (Sakuraba *et al.*, 2003). In addition, Rashid *et al.* characterized a novel phosphopentomutase (TK1777, COG1109), which is involved in the isomerization of the pentose. As *T. kodakaraensis* is not able to grow on (deoxy)nucleosides, the 'DERA pathway' of some Archaea probably functions in the anabolic direction like in Eukarya, but in contrast to Bacteria. The phyletic distribution of DERA shows a huge variety in Archaea that do contain this enzyme: for example some methanogens do not have it, but it is present in all the halophiles. This observation suggests that the pathway in halophiles could have a role in the synthesis of pentoses.

Adenosine 5'- monophosphate (AMP) metabolism pathway

Recently, a novel adenosine 5′-monophosphate (AMP) degrading pathway was discovered (Sato *et al.*, 2007). In *Thermococcus kodakaraensis,* and several other anaerobic euryarchaea, the first step of this pathway is cleaving AMP molecule with phosphate into adenine and ribose-1,5-phosphate by an AMP phosphorylase (TK0352/COG0213, Table 3.1). The ribose-1,5-phosphate moiety can subsequently be isomerized by ribose-1,5-bisphosphate isomerase (TK0185/COG1184) to ribulose-1,5-bisphophate. Finally, the RuBisCO type III (TK2290/COG1850) converts, with CO_2 and H_2O, ribulose-1,5-bisphosphate to two molecules of the central carbon metabolite 3-phosphoglycerate. The product 3-phosphoglycerate can be further metabolized to produces ATP. Thus, Archaea could use the pathway when energy levels are low. A second function would be CO_2 fixation; however, *T. kodakaraensis* uses the RuMP to generate riboses and therefore would lose the fixated carbon through formaldehyde (Sato *et al.*, 2007).

6-Deoxy-5-ketofructose-1-phosphate (DKFP) pathway

A novel erythrose 4-phosphate-independent chorismate pathway has been identified in *Methanocaldococcus jannaschii* (Tumbula *et al.*, 1997; White, 2004), and recently in *Methanococcus maripulus* (Porat *et al.*, 2006). The classical chorismate synthesis pathway starts with the condensation of erythrose-4-phosphate and phosphoenolpyruvate and the subsequent conversion to 3-dehydroquinate (DHQ), a precursor of chorsimate (Fig. 3.1). The first step is catalyzed by 3-deoxy-D-*arabino*-2-heptulosonate-7-phosphate (DAHP) synthase. Enzymes that are responsible for this conversion can be found in three different COGs (COG0722, COG3200, and COG2876, Table 3.1). Thus, proteins from different families convert the same reaction, which can be explained by the phenomenon that is called non-orthologous gene displacement (Koonin *et al.*, 1996). The enzymes catalyzing the second step are, however, all members of one distinct COG: DHQ synthase (COG0337). Because of the deviating pentose metabolism of the Archaea, erythrose-4-phosphate is not generated; rather, an alternative 6-deoxy-5-ketofructose-1-phosphate (DKFP) pathway has evolved that leads to the synthesis of 3-hydroquinate (DHQ) (Fig. 3.2). Recently, the biosynthetic route of DKFP in *M. jannaschii* has been described (White and Xu, 2006). MJ1585 (COG1830), an archaeal Class I fructose-1,6-bisphosphate aldolase (Siebers *et al.*, 2001), also catalyzes the transaldolase reaction of fructose-1-phosphate or fructose-1,6-bisphosphate with methylglyoxal to DKFP. Methylglyoxal can be generated chemically or enzymatically from the central metabolite glyceraldehyde-3-phosphate.

2-Amino-3,7-dideoxy-d -threo-hept-6-ulosonic (ADTH) synthase

The first step of the DKFP pathway is catalyzed by ADTH synthase converting DKFP to 2-amino-3,7-dideoxy-D-threo-hept-6-ulosonate. This biochemical transaldolase activity was measured in *M. jannaschii,* and the enzyme was identified (MJ0400, a paralog of MJ1585/COG1830) (White, 2004). In addition, a genetic study shows that the orthologous protein in *M. maripaludis* (MMP0686) is essential for growth without aromatic amino acids and the activity is inhibited during growth on aryl acids (Porat *et al.*, 2006).

3-Dehydroquinate synthase

The subsequent step is the synthesis of 3-dehydroquinate, which is catalyzed by MJ1249 (COG1465) (White, 2004). The DHQ synthase activity could also be detected in *M. maripaludis.* Unexpectedly, a *M. maripaludis* mutant, in which the gene encoding the homologous Mmp0006 was disrupted, could still grow without aromatic amino acids. Therefore, it was concluded that methanococci have an alternative activity for this step (Porat *et al.*, 2006).

The occurrence of the 3-dehydroquinate synthase in archaeal genomes correlates, although not perfectly, with the absence of the transketolase genes (see: NOPPP) and the classical DHQ synthase gene in methanogens, halophiles, *Archaeoglobus fulgidus* and *Cenarchaeum symbiosum* (Soderberg, 2005; Verhees *et al.*, 2003). It also correlates with the organisms that have one or more paralogs of fructose-1,6-bisphosphate aldolase (Porat *et al.*, 2006; Soderberg, 2005) and, furthermore, both DKFP pathway genes are often in proximity to genes involved in aromatic amino acid biosynthesis (Porat *et al.*, 2006). Thus, genetic, biochemical and bioinformatic evidence has revealed the existence of two archaeal pathways to generate DHQ, the classical chorismate pathway and the novel DKFP pathway. The DFKP pathway is probably not unique to Archaea as homologous genes have been found in bacterial genomes.

Concluding remarks: pentose anabolism

The 'anabolic' pentose pathways in Archaea is a combination of unique conversions and novel proteins, but it also consists of general features and orthologous proteins that can be found in all three domains of life. The 'anabolic' metabolism of pentoses reflects the diversity of archaeal metabolism with a 'conserved archaeal core' and with a 'variable shell' (Makarova *et al.*, 1999) and confirms that metabolic pathways in Archaea are ‘a playground of non-orthologous gene displacement’ (Koonin and Galperin, 2003).

Archaeal pentose catabolic metabolism

Introduction

While the hexose metabolism in Archaea has been studied extensively for many years, studies on their pentose utilization appeared only recently (Brouns *et al.*, 2006; Johnsen and Schonheit, 2004). In contrast to growth on hexose compounds, only a few archaeal species can metabolize C5 sugars, including some *Sulfolobus* spp. (Brouns *et al.,* 2006; Grogan, 1989) and several halophiles (Tindall, 1992). The most abundant pentoses in nature (L-arabinose, D-arabinose, D-xylose and D-ribose) have been reported to serve as sole carbon and energy source for these Archaea.

In Bacteria, yeast and fungi, pentose utilization has been studied extensively. The pentoses D-arabinose, D-ribose, D-xylose and L-arabinose can be metabolized in three different pathways. First, in Bacteria an isomerase, kinase and, if necessary, an epimerase can convert D-/L-arabinose and D-xylose into D-xylulose-5-phosphate (Fig. 3.3A), which is an intermediate of the NOPPP and phosphoketolase pathways. The phosphoketolase pathway (PKP) is a catabolic pathway found in several Bacteria (Biesterveld *et al.*, 1994) and yeasts (Evans and

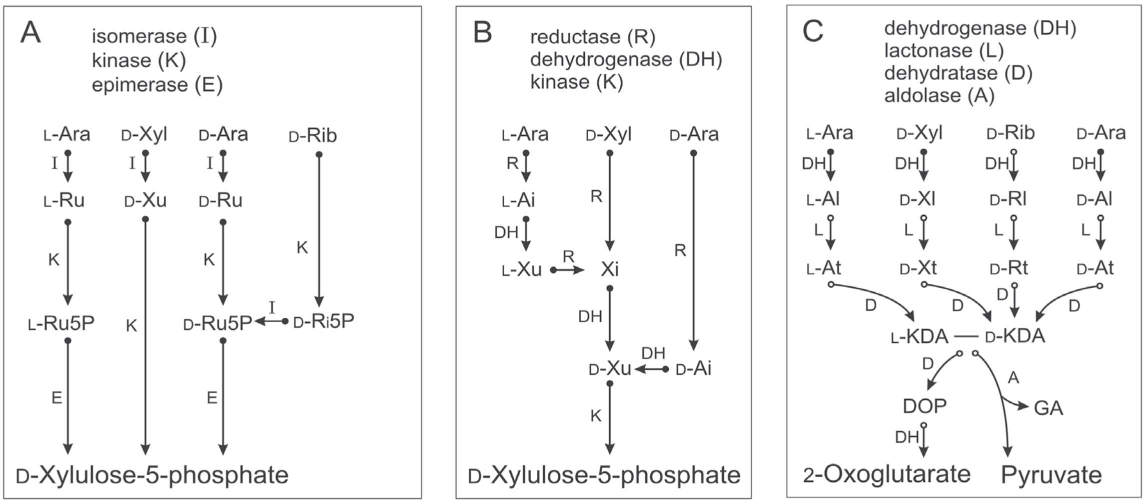


Figure 3.3 Pentose catabolism in Bacteria, Archaea and Eukarya. Schematic representation of three types of pentose degrading pathways (A, B and C). Arrows with an open or closed arrow tail represent enzymatic steps that are performed by unknown proteins or known proteins, respectively. Abbreviations: Ara, arabinose; Xyl, xylose; Ri(b), ribose; Ru, ribulose; Xu, xylulose; Ai, arabinitol; Xi, xylitol; Al, arabinonolactone; Xl, xylonolactone; Rl, ribonolactone; At, arabinonate; Xt, xylonate; Rt, ribonate; KDA, 2-keto-3-deoxy-arabinonate (also called 2-oxo-4,5-dihydroxypentanoate); DOP, 2,5-dioxopentanoate (also called 2-oxoglutarate semialdehyde); GA, glycolaldehyde. The figure was adapted from Brouns *et al.* (2006).

Table 3.2 Pentose catabolic enzymes in Archaea and their bacterial and eukaryal counterparts

Protein ID	Protein Name	EC[a]	COG[b]	ABE[c]	Characterized in Archaea[d]	Reference
Bacterial xylulose-5-phosphate pathway						
L-Arabinose						
AraA	L-Arabinose isomerase	5.3.1.4	COG2160	B		
AraB	L-Ribulokinase	2.7.1.16	COG1069	BE		
AraD	L-Ribulose-phosphate 4-epimerase	5.1.3.4	COG0235	ABE		
D-Xylose						
XylA	D-Xylose isomerase	5.3.1.5	COG2115	B		
XylB	D-Xylulokinase	2.7.1.17	COG1070	ABE		
D-Arabinose						
FucI	L-Fucose isomerase/D-Arabinose isomerase	5.3.1.25/ 5.3.1.3	COG2407	B		
DarK	D-Ribulokinase	2.7.1.47	COG1069	BE		
Rpe	Ribulose-phosphate 3-epimerase	5.1.3.1	COG0036	ABE		
D-Ribose						
RbsK	Ribokinase	2.7.1.15	COG0524	ABE		
RpiA	Ribose-5-phosphate isomerase	5.3.1.6	COG0120	ABE		
RpiB	Ribose-5-phosphate isomerase	5.3.1.6	COG0698	B		
Phosphoketolase pathway (PKP)						
Xfp	xylulose-5-phosphate/D-fructose 6-phosphate phosphoketolase	4.1.2.9/ 4.1.2.22	COG3957	BE		
General xylulose-5-phosphate pathway						
AlrA	Aldose reductase	1.1.1.21	COG0656	ABE		
LadA	L-Arabinitol 4-dehydrogenase	1.1.1.12	COG1063	ABE		
DcxR	L-Xylulose reductase	1.1.1.10	COG1028	ABE		
Xdh	D-Xylulose reductase	1.1.1.9	COG1063	ABE		
XylB	D-Xylulokinase	2.7.1.17	COG1070	ABE		
Prokaryotic pentose oxidation pathways						
D-Arabinose						
AraDH	D-Arabinose 1-dehydrogenase (NAD(P)$^+$)	1.1.1.117	COG1064	ABE	SSO1300	(Brouns *et al.*, 2006)
-	D-Arabinose 1-dehydrogenase	1.1.1.116	-	-		
-	D-Arabinonolactonase	3.1.1.30	COG3386	ABE		
AraD	D-Arabinonate dehydratase	4.2.1.5	COG4948	AB	SSO3124	(Brouns *et al.*, 2006)
KdaD	2-Dehydro-3-deoxy-D-arabinonate dehydratase	4.2.1.-	COG3970	AB	SSO3118	(Brouns *et al.*, 2006)
DopDH	2,5-Dioxovalerate dehydrogenase	1.2.1.26	COG1012	ABE	SSO3117	(Brouns *et al.*, 2006)
-	2-Dehydro-3-deoxy-D-pentonate aldolase	4.1.2.28	-	-		
D-Xylose						
Xdh	D-Xylose 1-dehydrogenase (NADP$^+$)	1.1.1.179	COG0673	ABE	rrnAC3034	(Johnsen and Schonheit, 2004)
Xyl1	D-Xylose 1-dehydrogenase	1.1.1.175	COG0673	ABE		
XylB	D-Xylose 1-dehydrogenase	1.1.1.175	COG1028	ABE		
-	D-Xylonolactonase	3.1.1.68	COG2220	ABE		
XylC	D-Xylonolactonase	3.1.1.68	COG3386	ABE		
-	D-Xylonate dehydratase	4.2.1.82	COG4948	AB		
XylD	D-Xylonate dehydratase	4.2.1.82	COG0129	ABE		
KdaD	2-Dehydro-3-deoxy-D -arabinonate dehydratase	4.2.1.-	COG3970	AB		

DopDH	2,5-Dioxovalerate dehydrogenase	1.2.1.26	COG1012	ABE
L-Arabinose				
AraA/AraE	L-Arabinose 1-dehydrogenase	1.1.1.46	COG0673	ABE
AraB/AraI	L-Arabinonolactonase	3.1.1.15	COG3386	ABE
AraC/AraB	L-Arabinonate dehydratase	4.2.1.25	COG0129	ABE
AraD/AraD	2-Dehydro-3-deoxy- L-arabinonate dehydratase	4.2.1.43	COG0329	AB
AraE/AraC	2,5-Dioxovalerate dehydrogenase	1.2.1.26	COG1012	ABE
Dahms pathway				
-	2-Dehydro-3-deoxy-D-pentonate aldolase	4.1.2.28	-	B
Transcriptional regulators involved in pentose metabolism				
AraR	Transcriptional repressor of the L-arabinose operon	-	COG1609	B
AraC	Transcriptional activator of the L-arabinose operon	-	COG2207	B
XylR	D-Xylose transcriptional activator	-	COG1609/ COG2207	B
XylR	D-Xylose operon repressor		COG1940	AB
RbsR	Transcriptional repressor of D-ribose operon	-	COG1609	B
Xyr1/XlnR	Xylanase regulator 1/ transcriptional activator XlnR	-	-	E

[a]EC, Enzyme Commission (Barthelmes *et al.*, 2007); [b]COG, Clusters of Orthologous Groups of Proteins (Tatusov *et al.*, 2003); [c]Domains are A: Archaea, B: Bacteria and E: Eukarya. [d]Locus id. The archaeal genome analysis is based on complete genome sequences of Table 3.1

Ratledge, 1984) and fungi. PKP converts hexoses, via the OPPP, or pentoses to xylulose-5-phosphate and subsequently split X5P with inorganic phosphate to glyceraldehyde-3-phosphate and acetyl-phosphate. The latter component can be further metabolized to acetate or ethanol. The enzyme xylulose-5-phosphate/D-fructose-6-phosphate phosphoketolase, which catalyzes this unique step in PKP, was characterized in *Bifidobacterium lactis* (xfp/COG3957) (Meile *et al.*, 2001). Orthologs of Xfp can be found in Eukarya; however, in Archaea no orthologs are present, which is in agreement with the fact that PKP has never been detected in this domain of life. The pentose D-ribose is, in contrast to the other pentoses, directly phosphorylated by a ribokinase (RbsK, COG0524) and metabolized through the NOPPP, in Bacteria (Hope *et al.*, 1986; Woodson and Devine, 1994). Genes encoding the 'bacterial xylulose-5-phosphate pathway' are generally clustered in the bacterial genomes. In *Escherichia coli* the *araBAD* operon is involved in L-arabinose degradation (Lee *et al.*, 1986). In the Gram-positive bacterium *Bacillus subtilis* the three encoding genes are in the *araABDLMNPQ-abfA* operon together with the genes encoding the arabinose transporter and an α–arabinofuranosidase, which cleaves L-arabinose monomers from arabinose oligomers (Mota *et al.*, 1999). The *xylAB* genes in *E. coli* encode an isomerase and kinase involved in D-xylose degradation (Rosenfeld *et al.*, 1984). While the uncommon D-arabinose is degraded by proteins encoded by the *darK-fucPIK* gene cluster of *E. coli* (Elsinghorst and Mortlock, 1994).

Second, pentoses can be converted to D-xylulose-5-phosphate by reductases and dehydrogenases (Fig. 3.3B). These pathways are found in fungi, mammals and yeast but also in some Bacteria (Chiang and Knight, 1960; Fossitt *et al.*, 1964; Wojtkiewicz *et al.*, 1988).

Third, the pentoses L-/D-arabinose, D-xylose and D-ribose can be metabolized to pyruvate and glycolaldehyde, or to 2-oxoglutarate (a tricarboxylic acid cycle intermediate) (Fig. 3.3C). The first common steps of the pentose conversion proceeds via a pentose dehydrogenase, a pentonolactonase, and a pentonic acid dehydratase. Then, 2-keto-3-deoxypentonic acid can be cleaved by a 2-dehydro-3-deoxy-D-pentonate aldolase into pyruvate and glycolaldehyde. This variant, also called the Dahms-pathway, has been observed in *Pseudomonas* and *Bradyrhizobium* strains (Dahms and Anderson, 1969; Palleroni and Doudoroff, 1957; Pedrosa and Zancan, 1974). Alternatively, 2-keto-3-deoxypentonic acid can be converted by a 2-keto-3-deoxypentonic acid dehydratase and a 2,5-dioxopentanoate dehydrogenase to generate 2-oxoglutarate. This pathway occurs in several aerobic Bacteria of the genera *Pseudomonas* (Dagley and Trudgill, 1965; Dahms, 1974; Weimberg, 1961; Weimberg and Doudoroff, 1955) *Rhizobium* (Duncan, 1979; Duncan and Fraenkel, 1979) and *Azospirillum* (Watanabe *et al.*, 2006a) and has recently been demonstrated in Archaea as well. *Sulfolobus solfataricus* can degrade D-arabinose to 2-oxoglutarate in four consecutive steps, catalyzed by the enzymes D-arabinose 1-dehydrogenase, D-arabinonate dehydratase, 2-dehydro-3-deoxy-D-arabinonate dehydratase and 2,5-dioxopentanoate (also called 2,5-dioxovalerate or α-ketoglutarate semialdehyde) dehydrogenase (Brouns *et al.*, 2006). Moreover, *Haloarcula marismortui* has been shown to utilize D-xylose as carbon and energy source; one of the enzymes involved D-xylose dehydrogenase, has been purified and characterized (Johnsen and Schonheit, 2004). The discussion below will focus on this class of prokaryotic pentose oxidation pathways (Fig. 3.3C) as these pathways are the only pentose degrading pathways in Archaea described to date.

Prokaryotic pentose oxidation pathways (PPOP)

The pentoses that can be metabolized via the third class of prokaryotic pentose oxidation pathways (Fig. 3.3C) are: D-arabinose in *Sulfolobus solfataricus* (Brouns *et al.*, 2006), L-arabinose in *Burkholderia thailandensis* (Moore *et al.*, 2004) and *Azospirillum brasilense* (Watanabe *et al.*, 2006a) and D-xylose in *Caulobacter crescentus* (Stephens *et al.*, 2007). In addition, this type of pathway has been predicted for D-xylose degradation in *Haloarcula marismortui* (Brouns *et al.*, 2006). Hence, this catabolic oxidation has only been described in *Sulfolobus* spp. and several halophiles in the domain Archaea, possibly reflecting the limited capacity of pentose utilization among Archaea. Some of the enzymes (Table 3.2) used in the pentose oxidation have been proposed to be similar to proteins that are possibly involved in hexaric acid and hydroxyproline catabolism in *Bacillus subtilis* and *Pseudomonas putida* (Brouns *et al.*, 2006). The first step is an oxidation of the pentose, after which, a possible pentonolactonase and dehydratase results in a 2-keto-3-deoxypentonic acid that can be dehydrated and oxidized to 2-oxoglutarate (Fig. 3.3C).

Pentose dehydrogenases

The oxidation of the pentose, D-/L-arabinose or D-xylose, is the first step in the prokaryotic pentose degradation. The dehydrogenation can be carried out by enzymes from different families. In *A. brasilense* the characterized L-arabinose 1-dehydrogenase (AraA) belongs to the COG0673 (Watanabe *et al.*, 2006a). AraA is 80% identical to AraE of *B. thailandensis*, which is essential for growth on L-arabinose (Moore *et al.*, 2004). Moreover, the distantly related and characterized D-xylose dehydrogenase from the halophilic *Haloarcula marismortui* belongs to the same COG0673 (Johnsen and Schonheit, 2004). It has been shown that the recombinant version of this D-xylose 1-dehydrogenase (rrnAC3034) prefers D-xylose, but can also oxidizes D-ribose and to a lesser extent D-glucose. The homotetrameric protein complex has NADP$^+$ as preferred electron acceptor, which is not uncommon for Archaea (Snijders *et al.*, 2006). AraA, in contrast, is a monomeric enzyme which catalyzes the oxidation of D-galactose as well (Watanabe *et al.*, 2006a). While it has been shown for *A. brasilense*, that it is capable of utilizing L-arabinose via the prokaryotic pentose oxidation pathway, a similar pathway has been proposed to be used by *H. marismortui* for converting D-xylose to 2-oxoglutarate, however, only a single enzymatic step has been characterized (Brouns *et al.*, 2006). The xylose 1-dehydrogenase activity has also been measured in the bacterium *Caulobacter crescentus* (Poindexter, 1964), and recently the genes of the complete D-xylose-degrading pathway were identified (Stephens *et al.*, 2007). The D-xylose dehydrogenase enzyme in *C. crescentus* XylB/CC0821 belongs to COG1028 and it uses NAD$^+$ as cofactor (Stephens *et al.*, 2007), while NADP$^+$ activity has been measured as well (Poindexter, 1964).

Finally, the archaeon *S. solfataricus* also degraded D-arabinose via the PPOP. The D-arabinose 1-dehydrogenase (AraDH; SSO1300) belongs to the COG1064 and forms a homotetrameric protein complex, with a clear cofactor preference for NADP$^+$ (Brouns *et al.*, 2006). Thus, the first oxidative step in the PPOP is used by various organisms to metabolize distinct pentoses that are catalyzed by different protein families.

Pentonolactonases

The oxidation of a pentose yields a pentonolactone which can be hydrolyzed by two different groups of proteins. First, COG3386, which includes the characterized D-gluconolactonase of *Zymomonas mobilis* (Kanagasundaram and Scopes, 1992), the recently identified L-arabinolactonase of *A. brasilense* (Watanabe *et al.*, 2006c) and the proposed D-xylonolactonase of *C. crescentus* (Stephens *et al.*, 2007). Second, the conversion can be catalyzed by an alternative version of the enzyme that belongs to COG2220. In *H. marismortui*, the D-xylonolactonase is probably member of the latter group (rrnAC3033) (Brouns *et al.*, 2006). In *Sulfolobus solfataricus,* the enzyme responsible for the conversion of D-arabinolactone into D-arabinonic acid might be a member of the former group (COG3386; SSO2705/SSO3041). However, it has been suggested that *S. solfataricus* does not need an enzyme for this conversion at all, since the spontaneous lactone hydrolysis reaction, possibly, occurs at non-limiting rates at high temperature (Brouns *et al.*, 2006).

Pentonate dehydratase dehydratases

The third step of the PPOP is the dehydration of a pentonic acid, which yields a 2-oxo-4,5-dihydroxypentanoate. In the bacterium *A. brasilense* the homodimer AraC is involved in the conversion of L-arabinonate into 2-keto-deoxy-L-arabinonate (Watanabe *et al.*, 2006c), and belongs to the dihydroxy-acid dehydratase (IlvD) and 6-phosphogluconate dehydratase (Edd) family (COG0129). IlvD/Edd is a huge protein family in which IlvD is involved in amino acid metabolism and Edd in the Entner–Doudoroff pathway. In *C. crescentus* four different proteins belong to this family and it has been proposed that CC0819/XylD is the D-xylonate dehydratase. Hence, the IlvD/Edd family has many different dehydratases that are not only involved in the amino acid and hexose catabolic pathways, but also in the distinct pentose catabolism.

In Archaea, unlike Bacteria, the D-arabinonate dehydratase from *S. solfataricus* (SSO3124) and the predicted D-xylonate dehydratase from *H. marismortui* (rrnAC3032) belong to the mandelate racemase/muconate lactonizing enzyme family and COG4948. SSO3124 forms a homooctameric complex, similarly as the homologous gluconate dehydratase (Kim and Lee, 2005). Interestingly, the enzyme that catalyzes the first step of the modified Entner–Doudoroff pathway in *S. solfataricus*, the glucose dehydrogenase (SSO3198), exhibits a high catalytic efficiency for D-xylose (Milburn *et al.*, 2006). In addition, in this archaeon the dihydroxy-acid dehydratase enzyme (COG0129, SSO3107) shows D-xylonate dehydratase activity as well (Kim and Lee, 2006). This suggests that SSO3107 and SSO3198 can play a role in the D-xylose catabolism of *S. solfataricus*, because the oxidation of D-xylose and D-arabinose yield the same intermediate (2-oxo-4(S),5-dihydroxypentanoate) and can be further metabolized to 2-oxoglutarate by the enzymes mentioned below.

2-Keto-3-deoxy-pentanoate dehydratases

The two different products that are the result of the pentonic acid dehydratases (1) 2-keto-3-deoxy-L-arabinonate and (2) 2-keto-3-deoxy-D-arabinonate (also called 2-keto-3-deoxy-D-xylonate), in which the chiral differences between the two pentoses at the C-2 and C-3 atoms have been eliminated by the pentonate dehydratase, are dehydrated again. Although two distinct protein families are responsible for catalyzing the dehydration of the two different compounds, the resulting product of this catalysis is the same: 2,5-dioxopentanoate (DOP, also called α-ketoglutaric semialdehyde). The first enzyme family, 2-keto-3-deoxy-L-arabinonate dehydratase (AraD), has first been identified in *A. brasilense* (COG0329) (Watanabe *et al.*, 2006a). The enzyme that generates α-ketoglutaric semialdehyde is homologous to the dihydrodipicolinate synthetase (DHDPS), but uses a unique mechanism. Whereas DHDSP catalyzes a C-C bond formation, AraD drives a dehydration reaction.

A second 2-keto-3-deoxy-arabinonate (Kda) dehydratase type that can convert 2-keto-3-deoxy-D-arabinonate into 2,5-dioxypentanoic acid has been identified in *S. solfataricus* (KdaD, SSO3118). The latter enzyme and the predicted Kda dehydratase (XylX, CC0823) in *C. crescentus* belong to the fumarylacetoacetate hydrolase family (COG3970), consisting of enzymes with diverse functions, such as decarboxylases and hydratases. The homologous rrnAC1339 is probably responsible for de dehydration reaction in the D-xylose catabolic pathway in *H. marismortui*. Besides the dehydration of Kda by the enzyme, family members in mammals including humans are capable of hydrolysing fumarylacetoacetate, the final step of phenylalanine and tyrosine degradation (Bateman *et al.*, 2001). Moreover, the C-terminal decarboxylation domain of the HpcE from *E. coli* is also member of this family (Tame *et al.*, 2002).

2,5-Dioxopentanoate dehydrogenase

The final step of the pentose, hexaric acid and L-hydroxyproline oxidation in the prokaryotic pathway concerns the dehydrogenation of the aldehyde 2,5-dioxopentanoate (DOP). The 2,5-dioxopentanoate dehydrogenase, which is also called α-ketoglutaric semialdehyde dehydrogenase, has been characterized in *S. solfataricus* (SSO3117, COG1012) (Brouns *et al.*, 2006). The archaeal homotetrameric enzyme prefers $NADP^+$ over NAD^+ and is a close homolog of the putative aldehyde dehydrogenases that could catalyze the final step of the D-xylose catabolism in *H. marismortui* (rrnAC3036) and in *C. crescentus* (XylA, CC0822). Surprisingly, the bacterium *A. brasilense* possesses three homologous isozymes that oxidize the final step of the PPOP, but are induced separately, when grown on L-arabinose, D-hexaric acids or L-hydroxyproline (see clustering of genes involved in pentose oxidation). These three DOP dehydrogenases are together with the archaeal dehydrogenases members of COG1012. Also in archaeal genome sequences many paralogs of the DOP dehydrogenases can be identified and therefore more archaeal enzymes might be capable of producing the TCA-cycle intermediate 2-oxoglutarate from DOP.

2-Dehydro-3-deoxy-D-pentonate aldolase

L-/D-KDA can also be cleaved into pyruvate and glycolaldehyde by the 2-dehydro-3-deoxy-D-pentonate aldolase (Dahms, 1974) (Fig. 3.3C). Although the activity was detected in *Pseudomonas* sp. and in *Bradyrhizobium* sp., the gene has never been identified and it is therefore not possible to predict if the so-called Dahms pathway is active in more species.

Clustering of genes involved in pentose oxidation

Genomic context analysis is a very powerful tool to predict protein function (Koonin *et al.*, 2001; Osterman and Overbeek, 2003; von Mering *et al.*, 2003). In prokaryotes conserved gene neighborhood or chromosomal gene clustering is a computational tool that can strongly suggest a certain function to a hypothetical protein, as turned out to be the case for the enzymes involved in the pentose, hexaric acids and L-hydroxyproline degradation pathways. In *S. solfataricus*, only the Kda dehydratase and DOP dehydrogenase genes of the identified enzymes of the D-arabinose pathway are located side by side. However, the comparative analysis of gene clusters in several aerobic proteobacteria that are member of the genera *Burkholderia*, *Rhizobium, Bradyrhizobium*, *Agrobacterium*, *Azospirillum*, and *Pseudomonas* correlates well with the ability of these organism to degrade pentoses, especially L-arabinose (Dagley and Trudgill, 1965; Duncan, 1979; Duncan and Fraenkel, 1979; Watanabe *et al.*, 2006a; Weimberg, 1961; Weimberg and Doudoroff, 1955). Based on genomic context analysis predictions, enzyme functions have been predicted for some missing links in the pathways (Brouns *et al.*, 2006). In the α-proteobacterium *C. crescentus* the D-xylose inducible promoter (Meisenzahl *et al.*, 1997) could be linked to D-xylose oxidation pathway, because the five downstream genes were assigned with enzymatical functions of the PPOP. Recently, almost all functional predictions have indeed been confirmed in this oligotrophic bacterium (Stephens *et al.*, 2007). The same D-xylose catabolic pathway was assigned to the archaeon *H. marismortui.* However, not all proteins are homologous to the enzymes of *C. crescentus* and the Kda dehydrates (rrnAC1339 and not rrnAC3039) is not in proximity to the other catabolic genes. Nevertheless, the enzymes can probably catalyze the same type of reactions, which has been shown for the D-xylose dehydrogenase enzyme (Johnsen and Schonheit, 2004).

The L-arabinose degradation gene cluster *araABCDEFGHI* in the pathogenic β-proteobacterium *Burkholderia thailandensis* consists of nine genes and was proposed to be responsible for the degradation to 2-oxoglutarate (Moore *et al.*, 2004). Disruption of the *araA*, *araC*, *araE*, and *araI* genes, which may encode a transcriptional regulator, DOP dehydrogenase, L-arabinose dehydrogenase and L-arabinolactonase, respectively, led to an L-arabinose negative phenotype. Interestingly, the *Azospirillum brasilense* has a gene cluster (*araBZYXADCR*) and a separately located DOP dehydrogenase (*araE*) for L-arabinose utilization. In this bacterium, these genes have been studied extensively and the L-arabinose degrading enzymes have been characterized (Watanabe *et al.*, 2006c). Strikingly, the gene clusters (*ycbC-ycbH*) with a homolog of the DOP dehydrogenase of *S. solfataricus* are most likely involved in hexaric acid degradation in *Bacillus* species (Sharma and Blumenthal, 1973). In addition, another gene cluster in *Pseudomonas putida* (PP1245–PP1249) is probably involved in the breakdown of L-hydroxyproline, which is a major constituent of collagen and plant cell wall proteins (Ramaswamy, 1984; Yoneya and Adams, 1961).

The degradation pathways of these diverse chemical compounds are, apparently, converging at the level of DOP. These mosaic pathways of orthologous and non-orthologous proteins involved in the catabolic reactions suggests that some of these enzymatic steps may have evolved by recruitment events (Jensen, 1976; Schmidt *et al.*, 2003), in which 2-oxoglutarate is the metabolic hub as the end product and may have been the driving force in the development of these pathways in aerobically respiring (TCA-cycle containing) Bacteria and Archaea.

Regulation of the pentose metabolism

In Bacteria and Eukarya, regulation of metabolism is executed at all levels. At the level of DNA by inducing or inhibiting transcription initiation, at the level of RNA by influencing transcription elongation and using attenuation mechanisms, and finally at the protein level through post-translational modification and allosteric regulation. Regulation at the protein level by allosteric regulation is certainly expected to be present in Archaea, because protein domains that are involved in allosteric regulation (*e.g.* ACT, RAM) are very common in Archaea (Aravind and Koonin, 1999; Ettema *et al.*, 2002). However, in the well-characterized archaeal glucose metabolic pathway, the Embden-Meyerhof-Parnas pathway (reviewed by (Siebers and Schonheit, 2005; van der Oost and Siebers, 2007; Verhees *et al.*, 2003)), the enzymes appear not to be regulated allosterically, but rather at the transcriptional level.

Regulation at the translational level (attenuation, anti-termination) has not been demonstrated to date in Archaea, but these mechanisms probably play an important role in archaeal biology as well. For instance, upstream of transport genes in *Thermoplasma* spp., thiamin riboswitches were predicted, which might be acquired by horizontal gene transfer (Rodionov *et al.*, 2002). In addition, an attenuation-like system has been proposed to regulate translation of the tryptophan regulator (TrpY) of *Methanobacterium thermoautotrophicus* (Xie and Reeve, 2005).

A relatively important site of regulating archaeal metabolism appears to be at the transcriptional level. The archaeal transcription initiation machinery is more similar to the eukaryal polymerase II system than to the bacterial machinery with its five subunit RNA polymerase (RNAP) and a series of sigma factors that interact with the Pribnow-box and the −35 region. The archaeal RNAP resembles the core of the eukaryal PolII system with (10–12 subunit RNAP, TATA-binding protein (TBP) and transcription factor B (TFB)) (Bell, 2005). Despite the similarity between the basal transcriptional machineries in Archaea and Eukarya, most archaeal transcriptional regulators more closely resemble their bacterial counterparts (Aravind and Koonin, 1999), but only a limited number of these regulators have been characterized (reviewed by (Bell, 2005; Geiduschek and Ouhammouch, 2005)). The archaeal transcriptional repressors often block the transcription initiation by competing with the transcription factors (TBP, TFB) or the RNAP for their respective binding site (*e.g.* Lrp-like regulators (Brinkman *et al.*, 2003)). Transcriptional activation has been shown in Archaea as well (Brinkman *et al.*, 2002; Ouhammouch *et al.*, 2003), little is known to date on specific interactions with the transcription initiation complex (TBP, TFB, RNAP).

The transcriptional regulator of the *mal* operon (TrmB) is the only transcriptional regulator involved in the archaeal carbohydrate metabolism that has been characterized. TrmB represses the transcription of the trehalose/maltose transport operon in *Thermococcus litoralis* (Lee *et al.*, 2003) and a separate maltodextrin ABC transporter in *P. furiosus* (Lee *et al.*, 2005). Recently, it was hypothesized that a TrmB homolog (Tgr) could be responsible for modulated gene expression of the archaeal glycolytic enzymes (van de Werken *et al.*, 2006); this has indeed been confirmed biochemically and genetically (Lee *et al.*, 2007; Kanai *et al.*, 2007). The transcriptional regulators of the archaeal pentose metabolism, however, are not known.

In Bacteria and Eukarya the oxidative branch of the pentose phosphate pathway is mainly regulated by the intracellular NADPH/NADP ratio. Glucose-6-phosphate dehydrogenase, for instance, is the rate-limiting step of the OPPP and this essentially irreversible conversion is allosterically regulated by the demand of the reducing power in the cell (Au *et al.*, 2000; Hansen *et al.*, 2002). The NOPPP is controlled, primarily, by the substrate availability (Berg *et al.*, 2002). As the OPPP is not likely to function in Archaea and several pathways are able to interconvert central metabolites with pentoses, different regulatory mechanisms are anticipated to occur in Archaea.

The bacterial pentose oxidation pathways are controlled mainly at the transcriptional level. In the Gram-positive bacterium *Bacillus subtilis* the AraR (COG1609, Table 3.2) negatively regulates the *araABDLMNPQ-abfA* operon and the *araE* and *araR* genes (Mota *et al.*, 1999) with L-arabinose as effector. In the Gram-negative organism *E. coli* the L-arabinose pathway is under control of AraC (COG2207), which activates in presence of L-arabinose the *araBAD* operon and the genes that transport L-arabinose (*araE* and *araF*) and its own product (Miyada *et al.*, 1984).

The XylR (COG1609/2207) of *E. coli* acts as an activator in the presence of D-xylose the transcription of the *xylAB* and the *xylFGHR* operons (Song and Park, 1997). In *B. subtilis* the XylR (COG1940) acts as a repressor of the *xylAB* operon by blocking its promoter (Dahl *et al.*, 1994). RbsR (COG1609) in both organisms represses the ribose operon, encoding the ribose kinase and the ribose transporter (Mauzy and Hermodson, 1992; Woodson and Devine, 1994).

The fungi *Hyprocrea jecorina* and *Aspergillus niger* are using the transcriptional activator Xyr1, XlnR respectively, to induce not only the transcription of the polysaccharide breakdown enzymes but also the D-xylose degradation genes (Hasper *et al.*, 2000; Stricker *et al.*, 2006). In Bacteria, the regulatory mechanism of the general xylulose-5-phosphate pathway is not fully known, but the enzymes involved in D-arabitol catabolism in *Aerobacter aerogenes* have been reported to be inducible (Wilson and Mortlock, 1973).

The prokaryotic pentose oxidation pathways (PPOP) in Bacteria seem to be modulated at the transcriptional level as well. Transcriptional profiling of *C. crescentus* showed up-regulation of the D-xylose gene cluster (2.8- to 11.6-fold) during growth on D-xylose vs. D-glucose (Hottes *et al.*, 2004), which confirms the D-xylose-dependent promoter (Meisenzahl *et al.*, 1997). In addition, reporter gene insertions showed that *araC* and *araE* gene expression of *Burkholderia thailandensis* was repressed during growth in D-glucose and was induced in L-arabinose media (Moore *et al.*, 2004). Finally, the three α-ketoglutarate semialdehyde dehydrogenase (KGSADH or DOP dehydrogenases) isozymes of *A. brasilense* are all inducible in different media: (KGSADH-I) is up-regulated grown on L-arabinose, (KGSADH-II) on D-glutarate/D-galactarate and (KGSADH-III) on hydroxy-L-proline (Watanabe *et al.*, 2006b; Watanabe *et al.*, 2007).

Although the PPOP gene clusters are regulated at the transcriptional level, no transcriptional regulators are characterized that are involved in the regulation of these genes. Most likely, however, the *araA* gene (COG0583) in *B. thailandensis* encodes a LysR-like positive regulator of the L-arabinose assimilation operon, since the *araA* knockout mutant was unable to grow on L-arabinose (Moore *et al.*, 2004).

Catabolite repression is the phenomenon that an organism will first fully consume the preferred carbohydrate (or the most efficient conversion, often glucose) and after a lag phase during which alternative enzymes are being produced, it will start degrading second-choice carbohydrates. This two-step utilization of carbohydrate mixes is known as 'catabolite repression'. In Bacteria and Eukarya catabolite repression is well studied, and important in pentose utilization. In Archaea, a catabolite repression-like system has been described in *Sulfolobus solfataricus* (Hoang *et al.*, 2004) (and reviewed by (Bini and Blum, 2001)). Recently, it has been shown that this catabolite repression-like system is also involved in regulation of the arabinose ABC-transporter (Lubelska *et al.*, 2006) and thus being an important mechanism in the pentose metabolism of this archaeon.

In *S. solfataricus* the complete D-arabinose pathway, including the arabinose transporter, was induced at transcriptome and proteome level grown on D-arabinose vs. D-glucose (Brouns *et al.*, 2006). This is similar with the results of the xylose dehydrogenase *H. marismortui*, which was induced during growth on xylose (Johnsen and Schonheit, 2004).

Recently, several functional genomic studies have been published (reviewed by van der Oost *et al.*, 2006). Moreover, integration of proteomics, transcriptomics and biochemistry has been successfully performed (Brouns *et al.*, 2006; Snijders *et al.*, 2006). The *Sulfolobus*

D-arabinose pathway analysis showed the added value of a complete transcriptome and proteome study. Four enzymatical functions were elucidated, but despite the prediction of a binding motif upstream of the genes involved in D-arabinose assimilation, no transcriptional regulator could be identified. Other archaeal functional studies, such as, the microarray analysis in *Pyrococcus furiosus*: reveal co-expression of genes encoding the glycolytic EMP enzymes, and key enzymes of amino acid biosynthesis and transketolase (clustered with AA biosynthesis), during growth on maltose vs. peptides (Schut *et al.*, 2003). However, a combined microarray/proteomics study in *Sulfolobus solfataricus* revealed almost no fluctuation (Snijders *et al.*, 2006).

Concluding remarks on the PPOP in Archaea and regulation of pentose metabolism

Only a few Archaea (Sulfolobales and halophiles) are able to grow on pentoses as sole carbon and energy source. These Archaea are using a prokaryotic pentose oxidation pathway that yields 2-oxoglutaric acid. This TCA intermediate can be completely oxidized by aerobic species. The PPOP enzymatic steps have probably been evolved by recruitment events, yet another example of a 'conserved housekeeping core' and 'variable metabolic shell' that allows adjusting the metabolic infrastructure to available substrates (Makarova *et al.*, 1999). The regulatory mechanisms are still unknown and therefore interesting future research areas, especially with the new transcriptome and proteome tools.

References

Ahmed, H., Ettema, T.J., Tjaden, B., Geerling, A.C., van der Oost, J., and Siebers, B. (2005). The semi-phosphorylative Entner–Doudoroff pathway in hyperthermophilic Archaea: a re-evaluation. Biochem. J. *390*, 529–540.

Aravind, L., and Koonin, E.V. (1999). DNA-binding proteins and evolution of transcription regulation in the Archaea. Nucleic Acids Res. *27*, 4658–4670.

Au, S.W., Gover, S., Lam, V.M., and Adams, M.J. (2000). Human glucose-6-phosphate dehydrogenase: the crystal structure reveals a structural NADP+ molecule and provides insights into enzyme deficiency. Structure *8*, 293–303.

Barthelmes, J., Ebeling, C., Chang, A., Schomburg, I., and Schomburg, D. (2007). BRENDA, AMENDA and FRENDA: the enzyme information system in 2007. Nucleic Acids Res. *35*, D511–514.

Bateman, R.L., Bhanumoorthy, P., Witte, J.F., McClard, R.W., Grompe, M., and Timm, D.E. (2001). Mechanistic inferences from the crystal structure of fumarylacetoacetate hydrolase with a bound phosphorus-based inhibitor. J. Biol. Chem. *276*, 15284–15291.

Bell, S.D. (2005). Archaeal transcriptional regulation – variation on a bacterial theme? Trends Microbiol. *13*, 262–265.

Berg, J.M., Tymoczko, J.L., and Stryer, L. (2002). Biochemistry, 5th edition (New York, W.H. Freeman and company).

Biesterveld, S., Kok, M.D., Dijkema, C., Zehnder, A.J., and Stams, A.J. (1994). D-xylose catabolism in *Bacteroides xylanolyticus* X5–1. Arch. Microbiol. *161*, 521–527.

Bini, E., and Blum, P. (2001). Archaeal catabolite repression: a gene regulatory paradigm. Adv. Appl. Microbiol. *50*, 339–366.

Brinkman, A.B., Bell, S.D., Lebbink, R.J., de Vos, W.M., and van der Oost, J. (2002). The *Sulfolobus solfataricus* Lrp-like protein LysM regulates lysine biosynthesis in response to lysine availability. J. Biol. Chem. *277*, 29537–29549.

Brinkman, A.B., Ettema, T.J., de Vos, W.M., and van der Oost, J. (2003). The Lrp family of transcriptional regulators. Mol. Microbiol. *48*, 287–294.

Brouns, S.J., Walther, J., Snijders, A.P., van de Werken, H.J., Willemen, H.L., Worm, P., de Vos, M.G., Andersson, A., Lundgren, M., Mazon, H.F*., et al.* (2006). Identification of the missing links in prokaryotic pentose oxidation pathways: evidence for enzyme recruitment. J. Biol. Chem. *281*, 27378–27388.

Chiang, C., and Knight, S.G. (1960). Metabolism of D-xylose by moulds. Nature *188*, 79–81.

Choquet, C.G., Richards, J.C., Patel, G.B., and Sprott, G.D. (1994). Ribose biosynthesis in methanogenic cacteria. Arch. Microbiol. *161*, 481–488.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. (2006). Toward automatic reconstruction of a highly resolved tree of life. Science *311*, 1283–1287.

Dagley, S., and Trudgill, P.W. (1965). The metabolism of galactarate, D-glucarate and various pentoses by species of *Pseudomonas*. Biochem. J. *95*, 48–58.

Dahl, M.K., Degenkolb, J., and Hillen, W. (1994). Transcription of the *xyl* operon is controlled in *Bacillus subtilis* by tandem overlapping operators spaced by four basepairs. J. Mol. Biol. *243*, 413–424.

Dahms, A.S. (1974). 3-Deoxy-D-pentulosonic acid aldolase and its role in a new pathway of D-xylose degradation. Biochem. Biophys. Res. Commun. *60*, 1433–1439.

Dahms, A.S., and Anderson, R.L. (1969). 2-keto-3-deoxyl-L-arabonate aldolase and its role in a new pathway of L-arabinose degradation. Biochem. Biophys. Res. Commun. *36*, 809–814.

De Rosa, M., Gambacorta, A., Nicolaus, B., Giardina, P., Poerio, E., and Buonocore, V. (1984). Glucose metabolism in the extreme thermoacidophilic archaebacterium *Sulfolobus solfataricus*. Biochem. J. *224*, 407–414.

Duncan, M.J. (1979). L-arabinose metabolism in rhizobia. J. Gen. Microbiol. *113*, 177–179.

Duncan, M.J., and Fraenkel, D.G. (1979). α-Ketoglutarate dehydrogenase mutant of *Rhizobium meliloti*. J. Bacteriol. *137*, 415–419.

Eisenreich, W., Schwarzkopf, B., and Bacher, A. (1991). Biosynthesis of nucleotides, flavins, and deazaflavins in *Methanobacterium thermoautotrophicum*. J. Biol. Chem. *266*, 9622–9631.

Ekiel, I., Smith, I.C., and Sprott, G.D. (1983). Biosynthetic pathways in *Methanospirillum hungatei* as determined by 13C nuclear magnetic resonance. J. Bacteriol. *156*, 316-326.

Elsinghorst, E.A., and Mortlock, R.P. (1994). Molecular cloning of the *Escherichia coli* B L-fucose-D-arabinose gene cluster. J. Bacteriol. *176*, 7223–7232.

Ettema, T.J., Brinkman, A.B., Tani, T.H., Rafferty, J.B., and Van Der Oost, J. (2002). A novel ligand-binding domain involved in regulation of amino acid metabolism in prokaryotes. J. Biol. Chem. *277*, 37464–37468.

Evans, C.T., and Ratledge, C. (1984). Induction of xylulose-5-phosphate phosphoketolase in a variety of yeasts grown on D-xylose – the key to efficient xylose metabolism. Arch. Microbiol. *139*, 48–52.

Fossitt, D., Mortlock, R.P., Anderson, R.L., and Wood, W.A. (1964). Pathways of L-arabitol and xylitol metabolism in *Aerobacter Aerogenes*. J. Biol. Chem. *239*, 2110–2115.

Geiduschek, E.P., and Ouhammouch, M. (2005). Archaeal transcription and its regulators. Mol. Microbiol. *56*, 1397–1407.

Grochowski, L.L., Xu, H., and White, R.H. (2005). Ribose-5-phosphate biosynthesis in *Methanocaldococcus jannaschii* occurs in the absence of a pentose-phosphate pathway. J. Bacteriol. *187*, 7382–7389.

Grogan, D.W. (1989). Phenotypic characterization of the archaebacterial genus *Sulfolobus*: comparison of five wild-type strains. J. Bacteriol. *171*, 6710–6719.

Hansen, T., Oehlmann, M., and Schonheit, P. (2001). Novel type of glucose-6-phosphate isomerase in the hyperthermophilic archaeon *Pyrococcus furiosus*. J. Bacteriol. *183*, 3428–3435.

Hansen, T., Schlichting, B., and Schonheit, P. (2002). Glucose-6-phosphate dehydrogenase from the hyperthermophilic bacterium *Thermotoga maritima*: expression of the g6pd gene and characterization of an extremely thermophilic enzyme. FEMS Microbiol. Lett. *216*, 249–253.

Hasper, A.A., Visser, J., and de Graaff, L.H. (2000). The *Aspergillus niger* transcriptional activator XlnR, which is involved in the degradation of the polysaccharides xylan and cellulose, also regulates D-xylose reductase gene expression. Mol. Microbiol. *36*, 193–200.

Hoang, V., Bini, E., Dixit, V., Drozda, M., and Blum, P. (2004). The role of *cis*-acting sequences governing catabolite repression control of *lacS* expression in the archaeon *Sulfolobus solfataricus*. Genetics *167*, 1563–1572.

Hope, J.N., Bell, A.W., Hermodson, M.A., and Groarke, J.M. (1986). Ribokinase from *Escherichia coli* K12. Nucleotide sequence and overexpression of the *rbsK* gene and purification of ribokinase. J. Biol. Chem. *261*, 7663–7668.

Hottes, A.K., Meewan, M., Yang, D., Arana, N., Romero, P., McAdams, H.H., and Stephens, C. (2004). Transcriptional profiling of *Caulobacter crescentus* during growth on complex and minimal media. J. Bacteriol. *186*, 1448–1461.

Ishikawa, K., Matsui, I., Payan, F., Cambillau, C., Ishida, H., Kawarabayasi, Y., Kikuchi, H., and Roussel, A. (2002). A hyperthermostable D-ribose-5-phosphate isomerase from *Pyrococcus horikoshii* characterization and three-dimensional structure. Structure *10*, 877–886.

Jensen, R.A. (1976). Enzyme recruitment in evolution of new function. Annu. Rev. Microbiol. *30*, 409–425.

Johnsen, U., and Schonheit, P. (2004). Novel xylose dehydrogenase in the halophilic archaeon *Haloarcula marismortui*. J. Bacteriol. *186*, 6198–6207.

Kanagasundaram, V., and Scopes, R. (1992). Isolation and characterization of the gene encoding gluconolactonase from *Zymomonas mobilis*. Biochim. Biophys. Acta *1171*, 198–200.

Kanai, T., Akerboom, J., Takedomi, S., van de Werken, H.J.G., Blombach, F., van der Oost, J., Murakami, T., Atomi, H., and Imanaka, T. (2007). A global transcriptional regulator in *Thermococcus kodakaraensis* controls the expression levels of both glycolytic and gluconeogenic enzyme-encoding genes. J. Biol. Chem. (in press).

Kemp, M.B. (1974). Hexose phosphate synthase from *Methylcoccus capsulatus* makes D-*arabino*-3-hexulose phosphate. Biochem. J. *139*, 129–134.

Kim, S., and Lee, S.B. (2005). Identification and characterization of *Sulfolobus solfataricus* D-gluconate dehydratase: a key enzyme in the non-phosphorylated Entner–Doudoroff pathway. Biochem. J. *387*, 271–280.

Kim, S., and Lee, S.B. (2006). Catalytic promiscuity in dihydroxy-acid dehydratase from the thermoacidophilic archaeon *Sulfolobus solfataricus*. J. Biochem. (Tokyo) *139*, 591–596.

Koonin, E.V., and Galperin, M.Y. (2003). Sequence–Evolution–Function: Computational Approaches in Comparative Genomics Kluwer Academic Publishers).

Koonin, E.V., Mushegian, A.R., and Bork, P. (1996). Non-orthologous gene displacement. Trends Genet. *12*, 334–336.

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

Lee, N., Gielow, W., Martin, R., Hamilton, E., and Fowler, A. (1986). The organization of the *araBAD* operon of Escherichia coli. Gene *47*, 231–244.

Lee, S.J., Engelmann, A., Horlacher, R., Qu, Q., Vierke, G., Hebbeln, C., Thomm, M., and Boos, W. (2003). TrmB, a sugar-specific transcriptional regulator of the trehalose/maltose ABC transporter from the hyperthermophilic archaeon *Thermococcus litoralis*. J. Biol. Chem. *278*, 983–990.

Lee, S.J., Moulakakis, C., Koning, S.M., Hausner, W., Thomm, M., and Boos, W. (2005). TrmB, a sugar sensing regulator of ABC transporter genes in *Pyrococcus furiosus* exhibits dual promoter specificity and is controlled by different inducers. Mol. Microbiol. *57*, 1797–1807.

Lee, S.J., Surma, M., Seitz, S., Hausner, W., Thomm, M., and Boos, W. (2007). Characterization of the TrmB-like protein, PF0124, a TGM-recognizing global transcriptional regulator of the hyperthermophilic archaeon *Pyrococcus furiosus*. Mol. Microbiol. *65*, 305–318.

Lubelska, J.M., Jonuscheit, M., Schleper, C., Albers, S.V., and Driessen, A.J. (2006). Regulation of expression of the arabinose and glucose transporter genes in the thermophilic archaeon *Sulfolobus solfataricus*. Extremophiles *10*, 383–391.

Makarova, K.S., Aravind, L., Galperin, M.Y., Grishin, N.V., Tatusov, R.L., Wolf, Y.I., and Koonin, E.V. (1999). Comparative genomics of the Archaea (Euryarchaeota): evolution of conserved protein families, the stable core, and the variable shell. Genome Res. *9*, 608–628.

Martinez-Cruz, L.A., Dreyer, M.K., Boisvert, D.C., Yokota, H., Martinez-Chantar, M.L., Kim, R., and Kim, S.H. (2002). Crystal structure of MJ1247 protein from *M. jannaschii* at 2.0 A resolution infers a molecular function of 3-hexulose-6-phosphate isomerase. Structure *10*, 195–204.

Mauzy, C.A., and Hermodson, M.A. (1992). Structural and functional analyses of the repressor, RbsR, of the ribose operon of *Escherichia coli*. Protein Sci. *1*, 831–842.

Meile, L., Rohr, L.M., Geissmann, T.A., Herensperger, M., and Teuber, M. (2001). Characterization of the D-xylulose 5-phosphate/D-fructose 6-phosphate phosphoketolase gene (*xfp*) from *Bifidobacterium lactis*.

J. Bacteriol. *183*, 2929–2936.

Meisenzahl, A.C., Shapiro, L., and Jenal, U. (1997). Isolation and characterization of a xylose-dependent promoter from *Caulobacter crescentus*. J. Bacteriol. *179*, 592–600.

Milburn, C.C., Lamble, H.J., Theodossis, A., Bull, S.D., Hough, D.W., Danson, M.J., and Taylor, G.L. (2006). The structural basis of substrate promiscuity in glucose dehydrogenase from the hyperthermophilic archaeon *Sulfolobus solfataricus*. J. Biol. Chem. *281*, 14796-14804.

Miyada, C.G., Stoltzfus, L., and Wilcox, G. (1984). Regulation of the *araC* gene of *Escherichia coli*: catabolite repression, autoregulation, and effect on *araBAD* expression. Proc. Natl. Acad. Sci. USA *81*, 4120–4124.

Moore, R.A., Reckseidler-Zenteno, S., Kim, H., Nierman, W., Yu, Y., Tuanyok, A., Warawa, J., DeShazer, D., and Woods, D.E. (2004). Contribution of gene loss to the pathogenic evolution of *Burkholderia pseudomallei* and *Burkholderia mallei*. Infect. Immun. *72*, 4172–4187.

Mota, L.J., Tavares, P., and Sa-Nogueira, I. (1999). Mode of action of AraR, the key regulator of L-arabinose metabolism in *Bacillus subtilis*. Mol. Microbiol. *33*, 476–489.

Ogushi, S., Ando, M., and Tsuru, D. (1984). Formaldehyde dehydrogenase from *Pseudomonas putida*: a zinc metalloenzyme. J. Biochem. (Tokyo) *96*, 1587–1591.

Orita, I., Sato, T., Yurimoto, H., Kato, N., Atomi, H., Imanaka, T., and Sakai, Y. (2006). The ribulose monophosphate pathway substitutes for the missing pentose phosphate pathway in the archaeon *Thermococcus kodakaraensis*. J. Bacteriol. *188*, 4698–4704.

Orita, I., Yurimoto, H., Hirai, R., Kawarabayasi, Y., Sakai, Y., and Kato, N. (2005). The archaeon *Pyrococcus horikoshii* possesses a bifunctional enzyme for formaldehyde fixation via the ribulose monophosphate pathway. J. Bacteriol. *187*, 3636-3642.

Osterman, A., and Overbeek, R. (2003). Missing genes in metabolic pathways: a comparative genomics approach. Curr. Opin. Chem. Biol. *7*, 238–251.

Ouhammouch, M., Dewhurst, R.E., Hausner, W., Thomm, M., and Geiduschek, E.P. (2003). Activation of archaeal transcription by recruitment of the TATA-binding protein. Proc. Natl. Acad. Sci. USA *100*, 5097–5102.

Pace, N.R. (1997). A molecular view of microbial diversity and the biosphere. Science *276*, 734–740.

Palleroni, N.J., and Doudoroff, M. (1957). Metabolism of carbohydrates by *Pseudomonas saccharophila*. III. Oxidation of D-arabinose. J. Bacteriol. *74*, 180–185.

Pedrosa, F.O., and Zancan, G.T. (1974). L-Arabinose metabolism in *Rhizobium japonicum*. J. Bacteriol. *119*, 336-338.

Poindexter, J.S. (1964). Biological Properties and Classification of the Caulobacter Group. Bacteriol. Rev. *28*, 231–295.

Porat, I., Sieprawska-Lupa, M., Teng, Q., Bohanon, F.J., White, R.H., and Whitman, W.B. (2006). Biochemical and genetic characterization of an early step in a novel pathway for the biosynthesis of aromatic amino acids and p-aminobenzoic acid in the archaeon *Methanococcus maripaludis*. Mol. Microbiol. *62*, 1117–1131.

Purwantini, E., and Daniels, L. (1998). Molecular analysis of the gene encoding F420-dependent glucose-6-phosphate dehydrogenase from *Mycobacterium smegmatis*. J. Bacteriol. *180*, 2212–2219.

Purwantini, E., Gillis, T.P., and Daniels, L. (1997). Presence of F420-dependent glucose-6-phosphate dehydrogenase in *Mycobacterium* and *Nocardia* species, but absence from *Streptomyces* and *Corynebacterium* species and methanogenic Archaea. FEMS Microbiol. Lett. *146*, 129–134.

Ramaswamy, S.G. (1984). Hydroxyproline 2-epimerase of *Pseudomonas*. Subunit structure and active site studies. J. Biol. Chem. *259*, 249–254.

Rashid, N., Imanaka, H., Fukui, T., Atomi, H., and Imanaka, T. (2004). Presence of a novel phosphopentomutase and a 2-deoxyribose 5-phosphate aldolase reveals a metabolic link between pentoses and central carbon metabolism in the hyperthermophilic archaeon *Thermococcus kodakaraensis*. J. Bacteriol. *186*, 4185–4191.

Rodionov, D.A., Vitreschak, A.G., Mironov, A.A., and Gelfand, M.S. (2002). Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. J. Biol. Chem. *277*, 48949–48959.

Rosenfeld, S.A., Stevis, P.E., and Ho, N.W. (1984). Cloning and characterization of the *xyl* genes from *Escherichia coli*. Mol. Gen. Genet. *194*, 410–415.

Roy, R., Mukund, S., Schut, G.J., Dunn, D.M., Weiss, R., and Adams, M.W. (1999). Purification and molecular characterization of the tungsten-containing formaldehyde ferredoxin oxidoreductase from the hyperthermophilic archaeon *Pyrococcus furiosus*: the third of a putative five-member tungstoenzyme family. J. Bacteriol. *181*, 1171–1180.

Sakuraba, H., Tsuge, H., Shimoya, I., Kawakami, R., Goda, S., Kawarabayasi, Y., Katunuma, N., Ago, H., Miyano, M., and Ohshima, T. (2003). The first crystal structure of archaeal aldolase. Unique tetrameric structure of 2-deoxy-D-ribose-5-phosphate aldolase from the hyperthermophilic Archaea Aeropyrum pernix. J. Biol. Chem. *278*, 10799–10806.

Sato, T., Atomi, H., and Imanaka, T. (2007). Archaeal type III RuBisCOs function in a pathway for AMP metabolism. Science *315*, 1003–1006.

Schleper, C., Jurgens, G., and Jonuscheit, M. (2005). Genomic studies of uncultivated Archaea. Nat. Rev. Microbiol. *3*, 479–488.

Schmidt, S., Sunyaev, S., Bork, P., and Dandekar, T. (2003). Metabolites: a helping hand for pathway evolution? Trends Biochem. Sci. *28*, 336-341.

Schonheit, P., and Schafer, T. (1995). Metabolism of Hyperthermophiles. World Journal of Microbiology & Biotechnology *11*, 26–57.

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

Sharma, B.S., and Blumenthal, H.J. (1973). Catabolism of D-glucaric acid to α-ketoglutarate in *Bacillus megaterium*. J. Bacteriol. *116*, 1346-1354.

Siebers, B., Brinkmann, H., Dorr, C., Tjaden, B., Lilie, H., van der Oost, J., and Verhees, C.H. (2001). Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase. J. Biol. Chem. *276*, 28710–28718.

Siebers, B., and Schonheit, P. (2005). Unusual pathways and enzymes of central carbohydrate metabolism in Archaea. Curr. Opin. Microbiol. *8*, 695–705.

Sinninghe Damste, J.S., Rijpstra, W.I., Hopmans, E.C., Prahl, F.G., Wakeham, S.G., and Schouten, S. (2002). Distribution of membrane lipids of planktonic Crenarchaeota in the Arabian Sea. Appl. Environ. Microbiol. *68*, 2997–3002.

Snel, B., Bork, P., and Huynen, M.A. (1999). Genome phylogeny based on gene content. Nat. Genet. *21*, 108–110.

Snijders, A.P., Walther, J., Peter, S., Kinnman, I., de Vos, M.G., van de Werken, H.J., Brouns, S.J., van der Oost, J., and Wright, P.C. (2006). Reconstruction of central carbon metabolism in *Sulfolobus solfataricus* using a two-dimensional gel electrophoresis map, stable isotope labelling and DNA microarray analysis. Proteomics *6*, 1518–1529.

Soderberg, T. (2005). Biosynthesis of ribose-5-phosphate and erythrose-4-phosphate in Archaea: a phylogenetic analysis of archaeal genomes. Archaea *1*, 347–352.

Soderberg, T., and Alver, R.C. (2004). Transaldolase of *Methanocaldococcus jannaschii*. Archaea *1*, 255–262.

Song, S., and Park, C. (1997). Organization and regulation of the D-xylose operons in *Escherichia coli* K-12: XylR acts as a transcriptional activator. J. Bacteriol. *179*, 7025–7032.

Stephens, C., Christen, B., Fuchs, T., Sundaram, V., Watanabe, K., and Jenal, U. (2007). Genetic Analysis of a Novel Pathway for D-Xylose Metabolism in *Caulobacter crescentus*. J. Bacteriol. *189*, 2181–2185.

Stricker, A.R., Grosstessner-Hain, K., Wurleitner, E., and Mach, R.L. (2006). Xyr1 (xylanase regulator 1) regulates both the hydrolytic enzyme system and D-xylose metabolism in *Hypocrea jecorina*. Eukaryot Cell *5*, 2128–2137.

Tame, J.R., Namba, K., Dodson, E.J., and Roper, D.I. (2002). The crystal structure of HpcE, a bifunctional decarboxylase/isomerase with a multifunctional fold. Biochemistry *41*, 2982–2989.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V., Krylov, D.M., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N*., et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

Tindall, B.J. (1992). The family *Halobacteriaceae*. In The prokaryotes. A handbook on the biology of Bacteria: ecophysiology, isolation, identification, applications (New York. N.Y., Springer-Verlag), pp. 768–808.

Tumbula, D.L., Teng, Q., Bartlett, M.G., and Whitman, W.B. (1997). Ribose biosynthesis and evidence for an alternative first step in the common aromatic amino acid pathway in *Methanococcus maripaludis*. J. Bacteriol. *179*, 6010–6013.

van de Werken, H.J., Verhees, C.H., Akerboom, J., de Vos, W.M., and van der Oost, J. (2006). Identification of a glycolytic regulon in the Archaea *Pyrococcus* and *Thermococcus*. FEMS Microbiol. Lett. *260*, 69–76.

van der Oost, J., and Siebers, B. (2007). The Glycolytic Pathways of Archaea – Evolution by Tinkering. In Archaea, R.A. Garret, and H.P. Klenk, eds. (Blackwell Publishing), pp. 247–259.

van der Oost, J., Walther, J., Brouns, S.J. J., van de Werken, H.J. G., Snijders, A.P. L., Wright, P.C., Andersson, A., Bernander, R., and de Vos, W.M. (2006). Functional Genomics of the Thermo-Acidophilic Archaeon

Sulfolobus Solfataricus. In Extremophiles – Methods in Microbiology F.A. Rainey, and A. Oren, eds. (Amsterdam, Elsevier/Academic Press), pp. 201–231.

Verhees, C.H., Huynen, M.A., Ward, D.E., Schiltz, E., de Vos, W.M., and van der Oost, J. (2001). The phosphoglucose isomerase from the hyperthermophilic archaeon *Pyrococcus furiosus* is a unique glycolytic enzyme that belongs to the cupin superfamily. J. Biol. Chem. *276*, 40926–40932.

Verhees, C.H., Kengen, S.W., Tuininga, J.E., Schut, G.J., Adams, M.W., De Vos, W.M., and Van Der Oost, J. (2003). The unique features of glycolytic pathways in Archaea. Biochem. J. *375*, 231–246.

von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., and Snel, B. (2003). STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. *31*, 258–261.

Vorholt, J.A., Marx, C.J., Lidstrom, M.E., and Thauer, R.K. (2000). Novel formaldehyde-activating enzyme in *Methylobacterium extorquens* AM1 required for growth on methanol. J. Bacteriol. *182*, 6645–6650.

Watanabe, S., Kodaki, T., and Makino, K. (2006a). Cloning, expression, and characterization of bacterial L-arabinose 1-dehydrogenase involved in an alternative pathway of L-arabinose metabolism. J. Biol. Chem. *281*, 2612–2623.

Watanabe, S., Kodaki, T., and Makino, K. (2006b). A novel α-ketoglutaric semialdehyde dehydrogenase: evolutionary insight into an alternative pathway of bacterial L-arabinose metabolism. J. Biol. Chem. *281*, 28876-28888.

Watanabe, S., Shimada, N., Tajima, K., Kodaki, T., and Makino, K. (2006c). Identification and characterization of L-arabonate dehydratase, L-2-keto-3-deoxyarabonate dehydratase, and L-arabinolactonase involved in an alternative pathway of L-arabinose metabolism. Novel evolutionary insight into sugar metabolism. J. Biol. Chem. *281*, 33521–33536.

Watanabe, S., Yamada, M., Ohtsu, I., and Makino, K. (2007). α-Ketoglutaric semialdehyde dehydrogenase Isozymes involved in metabolic pathways of D-glucarate, D-galactarate, and hydroxy-L-proline: molecular and metabolic convergent evolution. J. Biol. Chem. *282*, 6685–6695.

Weimberg, R. (1961). Pentose oxidation by *Pseudomonas fragi*. J. Biol. Chem. *236*, 629–635.

Weimberg, R., and Doudoroff, M. (1955). The oxidation of L-arabinose by *Pseudomonas saccharophila*. J. Biol. Chem. *217*, 607–624.

White, R.H. (2004). L-Aspartate semialdehyde and a 6-deoxy-5-ketohexose 1-phosphate are the precursors to the aromatic amino acids in *Methanocaldococcus jannaschii*. Biochemistry *43*, 7618–7627.

White, R.H., and Xu, H. (2006). Methylglyoxal is an intermediate in the biosynthesis of 6-deoxy-5-ketofructose-1-phosphate: a precursor for aromatic amino acid biosynthesis in *Methanocaldococcus jannaschii*. Biochemistry *45*, 12366-12379.

Wilson, B.L., and Mortlock, R.P. (1973). Regulation of D-xylose and D-arabitol catabolism by *Aerobacter aerogenes*. J. Bacteriol. *113*, 1404–1411.

Woese, C.R., and Fox, G.E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc. Natl. Acad. Sci. USA *74*, 5088–5090.

Woese, C.R., Kandler, O., and Wheelis, M.L. (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc. Natl. Acad. Sci. USA *87*, 4576–4579.

Wojtkiewicz, B., Szmidzinski, R., Jezierska, A., and Cocito, C. (1988). Identification of a salvage pathway for D-arabinose in *Mycobacterium smegmatis*. Eur. J. Biochem. *172*, 197–203.

Woodson, K., and Devine, K.M. (1994). Analysis of a ribose transport operon from *Bacillus subtilis*. Microbiology *140 (Pt 8)*, 1829–1838.

Xavier, K.B., da Costa, M.S., and Santos, H. (2000). Demonstration of a novel glycolytic pathway in the hyperthermophilic archaeon *Thermococcus zilligii* by 13C-labeling experiments and nuclear magnetic resonance analysis. J. Bacteriol. *182*, 4632–4636.

Xie, Y., and Reeve, J.N. (2005). Regulation of tryptophan operon expression in the archaeon *Methanothermobacter thermautotrophicus*. J. Bacteriol. *187*, 6419–6429.

Yoneya, T., and Adams, E. (1961). Hydroxyproline metabolism. V. Inducible allohydroxy-d-proline oxidase of Pseudomonas. J. Biol. Chem. 236, 3272–3279.

Yu, J.P., Ladapo, J., and Whitman, W.B. (1994). Pathway of glycogen metabolism in Methanococcus maripaludis. J. Bacteriol. 176, 325–332.

Chapter 4

Identification of a glycolytic regulon in the archaea *Pyrococcus* and *Thermococcus*

van de Werken, H. J. G., Verhees, C. H., Akerboom, J., de Vos, W. M., and van der Oost, J. (2006). FEMS Microbiol Lett *260*, 69-76.

Abstract

The glycolytic pathway of the hyperthermophilic archaea that belong to the order *Thermococcales* (*Pyrococcus, Thermococcus and Palaeococcus)* differs significantly from the canonical Embden-Meyerhof pathway in bacteria and eukarya. This archaeal glycolysis variant consists of several novel enzymes, some of which catalyze unique conversions. Moreover, the enzymes appear not to be regulated allosterically, but rather at transcriptional level. To elucidate details of the gene expression control, the transcription initiation sites of the glycolytic genes in *Pyrococcus furiosus* have been mapped by primer extension analysis and the obtained promoter sequences have been compared to upstream regions of non-glycolytic genes. Apart from consensus sequences for the general transcription factors (TATA-box and BRE) this analysis revealed the presence of a potential transcription factor binding site (TATCAC-N_5-GTGATA) in glycolytic and starch utilizing promoters of *P. furiosus* and several thermococcal species. The absence of this inverted repeat in *P. abyssi* and *P. horikoshii* probably reflects that their reduced catabolic capacity does not require this regulatory system. Moreover, this phyletic pattern revealed a TrmB-like regulator (PF0124 and TK1769) which may be is involved in recognizing the TGM. This *Thermococcales* glycolytic regulon, with more than 20 genes, is the largest regulon that has yet been described for Archaea.

Introduction

A combination of metabolic, biochemical and genetic approaches have shown that the glycolysis in the hyperthermophilic archaea that belong to the order *Thermococcales* (*Pyrococcus* spp., *Thermococcus* spp. and *Palaeococcus* spp.) differs from the classical bacterial and eukaryal pathway because of different conversions, novel enzymes, and a distinct control (reviewed by (Verhees *et al.*, 2003)). In the classical Embden-Meyerhof pathway, the irreversible phosphorylation reactions catalyzed by hexokinase, phosphofructokinase and pyruvate kinase are allosterically regulated. In *P. furiosus,* however, the ADP-dependent glucokinase, ADP-dependent phosphofructokinase and pyruvate kinase are not controlled by any of the usual effector molecules (Tuininga, 2004; Tuininga *et al.*, 1999; Verhees *et al.*, 2002). Another potential regulatory site of archaeal glycolysis may be the apparent irreversible oxidation of glyceraldehyde-3-phosphate by an archaeal-type ferredoxin-dependent oxidoreductase (GAPOR). Similarly, no regulation has been reported at enzyme level, but rather at the level of gene expression (van der Oost *et al.*, 1998). Other studies showed that this is a general trend: glycolytic enzymes in *P. furiosus* are mainly, if not completely, regulated at transcriptional level (Siebers *et al.*, 2001; Verhees *et al.*, 2001). This has been confirmed by recent DNA microarray analyses that demonstrated the modulated expression of the glycolytic genes in *P. furiosus* (Schut *et al.*, 2003; Schut *et al.*, 2001; Weinberg *et al.*, 2005).

In bacteria and eukaryotes, glycolysis can be positively or negatively regulated at the level of gene expression. In gram-positive bacteria, the catabolite control protein (CcpA) is a repressor of many catabolic operons, but is also a transcriptional activator of glycolytic operons including genes encoding a phosphofructokinase, a pyruvate kinase and a lactate dehydrogenase (Luesink *et al.*, 1998; van den Bogaard *et al.*, 2000). In bacterium *E. coli* the catabolite repressor-activator protein (Cra) , formerly known as fructose repressor protein (FruR), negatively regulates transcription of genes encoding glycolytic enzymes, and positively regulates transcription of genes encoding gluconeogenic enzymes (Ramseier *et al.*, 1995). In yeast, a DNA-binding protein (GCR1) strongly reduces the transcription levels of most genes that encode glycolytic enzymes (Baker, 1991). To our knowledge however, no transcriptional regulator modulates the gene expression of all glycolytic genes and no homologs of the aforementioned regulators have been identified in archaeal genomes.

The basal transcription machinery in Archaea represents a simplified version of the RNA polymerase (RNAP) II transcription apparatus in Eukarya. The archaeal RNAP consists of 12-subunits and requires two general transcription factors for initiating transcription: Transcription Factor B (TFB) and TATA-binding protein (TBP). TFB and TBP bind to the Transcription Factor B-responsive element (BRE) and the TATA-box, respectively, and mediate the recognition of the archaeal promoter (Bell *et al.*, 2001).

Despite the similarity between the basal transcriptional machineries in Archaea and

Eukarya, most archaeal transcriptional regulators resemble bacterial counterparts (Aravind and Koonin, 1999). Only a limited number of the archaeal regulators have been characterized, mainly involved in metal homeostasis and amino acid metabolism (reviewed by (Geiduschek and Ouhammouch, 2005)). Recently, the first archaeal transcriptional regulator involved in sugar metabolism has been characterized in *Thermococcus litoralis*. This transcriptional regulator of the *mal* operon (TrmB) represses the transcription of the trehalose/maltose transport operon (Lee *et al.*, 2003). In *P. furiosus* the TrmB, which has probably been horizontal transferred between the two organisms (Diruggiero *et al.*, 2000), is identical and controls also the maltodextrin ABC-transporter (Lee *et al.*, 2005). However, no transcriptional regulators are currently known to be responsible for modulated gene expression of the archaeal glycolytic enzymes.

This study was initiated to determine the transcription initiation sites of the glycolytic genes in *P. furiosus* and to compare their promoter structures to identify functionally important elements. Apart from a BRE and TATA-box consensus, a conserved inverted repeat is identified in the promoter sequences of all glycolytic genes and several other genes involved in sugar metabolism in *P. furiosus* and *T. kodakaraensis*. The physiological implications of this potential transcription factor binding site are discussed, and integrated with recently reported experimental analyses of sugar metabolism in *Thermococcales*.

Materials and Methods

Organism and growth conditions.

P. furiosus (DSM 3638) was grown in a chemically defined medium, as previously described (Kengen *et al.*, 1993) with the only difference that yeast extract was replaced by the individual amino acids (0.25 mM final concentration). Maltose (10 mM) or pyruvate (40 mM) was used as the primary carbon source.

Table 4.1 5'-(IRD800)-labeled antisense oligonucleotides.

Gene Name	Nucleotide sequence	Target residues[a]
glk	5'-TGTCCAAGTATTTTATAGCGTCG-3'	102-124
pgi	5'-CTTTCCATGCCCTTTCATCAAC-3'	103-124
pfk	5'-ATTTTATCGGGACCAAATTCC-3'	102-122
fba	5'-CAAAGTCCGTAGGGCCGTGC-3'	99-118
tpi	5'-AATTGTTACACCTGTTTCTTTGTAC-3'	102-126
gor	5'-ATGTCCTTAGTTCATTGTGTCTC-3'	102-124
pyk	5'-ATTCTTGCAACATTCATCCCCG-3'	89-110
pps	5'-TGGTGGAACTGGAATTCCAGC-3'	97-117

[a] The numbers indicate the position of the nucleotides downstream the translation start site.

Transcript analysis.

RNA was isolated from maltose and pyruvate grown *P. furiosus* cells as previously described (Ward *et al.*, 2000). The transcription start sites were determined with fluorescent (IRD800)-labelled antisense oligonucleotides (Table 4.1). Primer extension reactions were performed using the Reverse Transcription System (Promega), according to the instructions of the manufacturer, with the following modifications: Hybridization of total RNA (15 µg) and oligonucleotide (5 pmol) was performed at 68 ºC for 10 min after which the sample was cooled to room temperature. The primer extension reaction (20 µL final volume) was started by addition of dNTPs (1 mM), $MgCl_2$ (5 mM), RNase inhibitor (RNAsin; 20 U) and avian myeloblastosis virus (AMV)-reverse transcriptase (22.5 U). After incubating for 30 min at 45 ºC the reaction volume was diluted to 50 µL with 10 mM Tris/HCl (pH 8.5) after which 1 µL of RNase A (5 mg mL^{-1}) was added. The mixture was subsequently incubated at 37 ºC for 10 minutes. Produced cDNA was precipitated with ethanol and dissolved in 3 µL loading buffer. 1 µL of sample was applied to a sequencing gel in parallel with sequencing reactions using the same oligonucleotide.

Computational analysis

The prokaryotic nucleotide sequences and annotations were downloaded from the National Center for Biotechnology Information References Sequences (RefSeq) (Pruitt *et al.*, 2005) and GenBank (Benson *et al.*, 2005) and the intergenic regions were extracted.

The BRE and the TATA-box in *P. furiosus* were identified, in a window of 14 nucleotides, using 27 nucleotide sequences from -42 to -16 before the experimentally determined transcription start sites. The nucleotide sequences were analyzed using the Gibbs Recursive Sampler algorithm (Thompson *et al.*, 2003). The glycolytic genes from the four completely sequenced *Thermococcales* genomes (*P. furiosus* (Accession number: NC_003413), *P. horikoshii* (NC_000961), *P. abyssi* (NC_000868) and *T. kodakaraensis* (NC_006624)) were used to identify the Thermococcales-Glycolytic-Motif (TGM).

The complete intergenic regions of *P. furiosus* and *T. kodakaraensis* were scanned with the TGM-matrix using the site search method of TFBS modules (Lenhard and Wasserman, 2002). The upstream sequences of the glycolytic genes in both species had a minimal value of 83%, which is used as cut-off value. Furthermore, all prokaryotes sequences were scanned with the TGM-matrix of *T. kodakaraensis* to determine additional occurrences of this putative cis-acting motif. The bi-directional best hit criterion (BLASTP (Altschul *et al.*, 1997), E-value < 1 x 10^{-5}) was used to identify orthologous proteins of *P. furiosus* in the predicted proteomes of *P. abyssi*, *P. horikoshii* and *T. kodakaraensis*.

Results and discussion

Mapping transcription start sites and promoter elements

The genes encoding the enzymes of the glycolytic pathway in *P. furiosus* have been identified by sequence analysis, or by determination of the N-terminus of the purified enzymes (Verhees *et al.*, 2003) (Supplementary Table S4.1). Based on experimental data, all the characterized

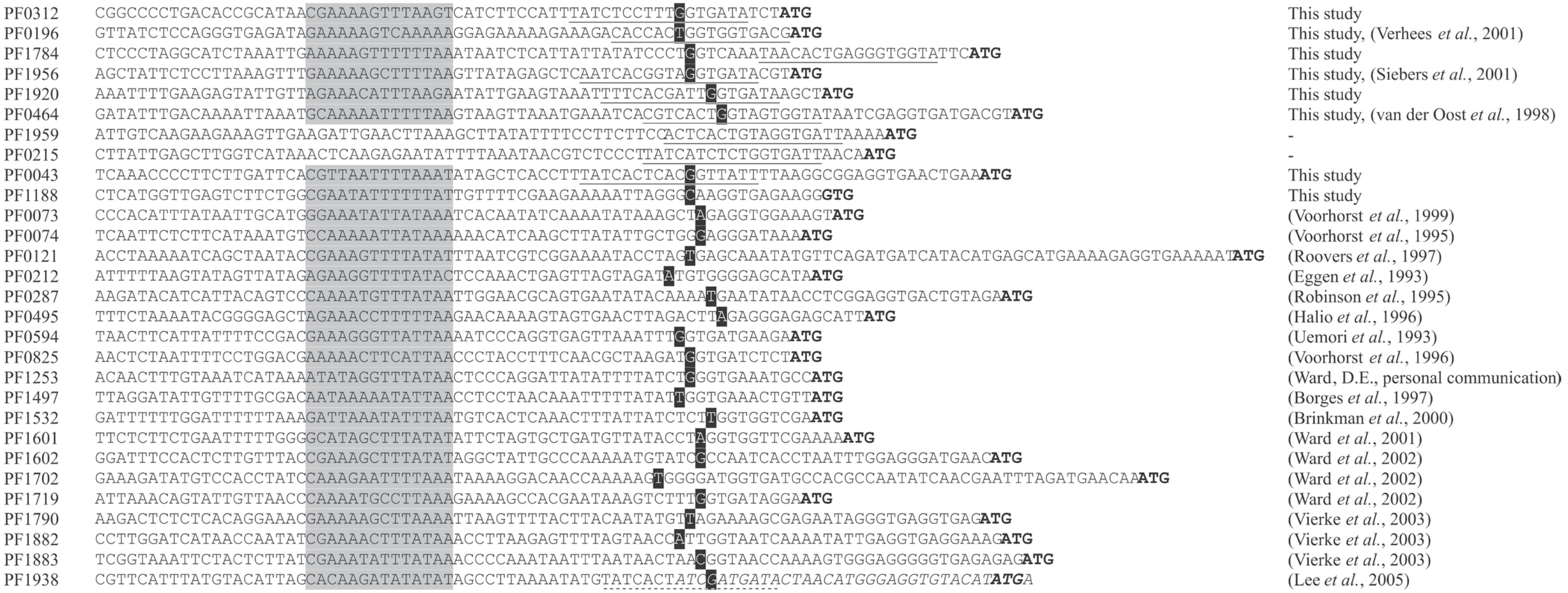

PF0312 CGGCCCCTGACACCGCATAACGAAAAGTTTAAGTCATCTTCCATTTATCTCCTTTGGTGATATCTATG This study
PF0196 GTTATCTCCAGGGTGAGATAGAAAAAGTCAAAAAGGAGAAAAAGAAAGACACCACTGGTGGTGACGATG This study, (Verhees et al., 2001)
PF1784 CTCCCTAGGCATCTAAATTGAAAAAGTTTTTAAATAATCTCATTATTATATCCCTGGTCAAATAACACTGAGGGTGGTATTCATG This study
PF1956 AGCTATTCTCCTTAAAGTTTGAAAAAGCTTTTAAGTTATAGAGCTCAATCACGGTAGGTGATACGTATG This study, (Siebers et al., 2001)
PF1920 AAATTTTGAAGAGTATTGTTAGAAACATTTAAGAATATTGAAGTAAATTTTCACGATTGGTGATAAGCTATG This study
PF0464 GATATTTGACAAAATTAAATGCAAAAATTTTTAAGTAAGTTAAATGAAATCACGTCACTGGTAGTGGTATAATCGAGGTGATGACGTATG This study, (van der Oost et al., 1998)
PF1959 ATTGTCAAGAAGAAAGTTGAAGATTGAACTTAAAGCTTATATTTTCCTTCTTCCACTCACTGTAGGTGATTAAAAATG -
PF0215 CTTATTGAGCTTGGTCATAAACTCAAGAGAATATTTTAAATAACGTCTCCCTTATCATCTCTGGTGATTAACAATG -
PF0043 TCAAACCCCTTCTTGATTCACGTTAATTTTAAATATAGCTCACCTTTATCACTCACGGTTATTTTAAGGCGGAGGTGAACTGAAATG This study
PF1188 CTCATGGTTGAGTCTTCTGGCGAATATTTTTTATTGTTTTCGAAGAAAAATTAGGGCAAGGTGAGAAGGGTG This study
PF0073 CCCACATTTATAATTGCATGGGAAATATTATAAATCACAATATCAAAATATAAAGCTAGAGGTGGAAAGTATG (Voorhorst et al., 1999)
PF0074 TCAATTCTCTTCATAAATGTCCAAAAATTATAAAAAACATCAAGCTTATATTGCTGGGAGGGATAAAATG (Voorhorst et al., 1995)
PF0121 ACCTAAAAATCAGCTAATACCGAAAGTTTTATATTTAATCGTCGGAAAATACCTAGTGAGCAAATATGTTCAGATGATCATACATGAGCATGAAAAGAGGTGAAAAATATG (Roovers et al., 1997)
PF0212 ATTTTTAAGTATAGTTATAGAGAAGGTTTTATACTCCAAACTGAGTTAGTAGATATGTGGGGAGCATAATG (Eggen et al., 1993)
PF0287 AAGATACATCATTACAGTCCCAAAATGTTTATAATTGGAACGCAGTGAATATACAAAATGAATATAACCTCGGAGGTGACTGTAGAATG (Robinson et al., 1995)
PF0495 TTTCTAAAATACGGGGAGCTAGAAACCTTTTTAAGAACAAAAGTAGTGAACTTAGACTTAGAGGGAGAGCATTATG (Halio et al., 1996)
PF0594 TAACTTCATTATTTTCCGACGAAAGGGTTATTAAAATCCCAGGTGAGTTAAATTTGGTGATGAAGAATG (Uemori et al., 1993)
PF0825 AACTCTAATTTTCCTGGACGAAAAACTTCATTAACCCTACCTTTCAACGCTAAGATGGTGATCTCTATG (Voorhorst et al., 1996)
PF1253 ACAACTTTGTAAATCATAAAATATAGGTTTATAACTCCCAGGATTATATTTTATCTGGGTGAAATGCCATG (Ward, D.E., personal communication)
PF1497 TTAGGATATTGTTTTGCGACAATAAAAATATTAACCTCCTAACAAATTTTTATATTGGTGAAACTGTTATG (Borges et al., 1997)
PF1532 GATTTTTTGGATTTTTTAAAGATTAAATATTTAATGTCACTCAAACTTTATTATCTCTTGGTGGTCGAATG (Brinkman et al., 2000)
PF1601 TTCTCTTCTGAATTTTTGGGGCATAGCTTTATATATTCTAGTGCTGATGTTATACCTAGGTGGTTCGAAAAATG (Ward et al., 2001)
PF1602 GGATTTCCACTCTTGTTTACCGAAAGCTTTATATAGGCTATTGCCCAAAAATGTATCGCCAATCACCTAATTTGGAGGGATGAACATG (Ward et al., 2002)
PF1702 GAAAGATATGTCCACCTATCCAAAGAATTTTAAATAAAAGGACAACCAAAAAGTGGGGATGGTGATGCCACGCCAATATCAACGAATTTAGATGAACAAATG (Ward et al., 2002)
PF1719 ATTAAACAGTATTGTTAACCCAAAATGCCTTAAAGAAAAGCCACGAATAAAGTCTTTGGTGATAGGAATG (Ward et al., 2002)
PF1790 AAGACTCTCTCACAGGAAACGAAAAAGCTTAAAATTAAGTTTTACTTACAATATGTTAGAAAAGCGAGAATAGGGTGAGGTGAGATG (Vierke et al., 2003)
PF1882 CCTTGGATCATAACCAATATCGAAAACTTTATAAACCTTAAGAGTTTTAGTAACCATTGGTAATCAAAATATTGAGGTGAGGAAAGATG (Vierke et al., 2003)
PF1883 TCGGTAAATTCTACTCTTATCGAAATATTTATAAACCCCAAATAATTTAATAACTAACGGTAACCAAAAGTGGGAGGGGGTGAGAGAGATG (Vierke et al., 2003)
PF1938 CGTTCATTTATGTACATTAGCACAAGATATATATAGCCTTAAAATATGTATCACTATCGATGATACTAACATGGGAGGTGTACATATGA (Lee et al., 2005)

Figure 4.1 Multiple alignment of the promoter sequences of *P. furiosus* with gene and locus names.Putative Transcription Factor B-responsive element (BRE) and TATA-box are highlighted in light gray. The identified Thermococcales-Glycolytic-Motifs (TGM) upstream of the genes encoding the glycolytic enzymes are underlined and the translation initiation codons are given in bold. Determined transcription start sites are indicated in black boxes. The region protected by TrmB determined by footprint analysis (Lee *et al.*, 2005) is indicated by italic letters and the predicted TGM in the promoter sequence of PF1938 is indicated by a dashed underline. The second promoter of PF1602 and PF1702 were used to align the BRE and TATA-box.PF0312, ADP-dependent glucokinase; PF0196, Phosphoglucose isomerase; PF1784, ADP-dependent phosphofructokinase; PF1956, Fructose-1,6-bisphosphate aldolase; PF1920, Triosephosphate isomerase; PF0464, Glyceraldehyde-3-phosphate: ferredoxin oxidoreductase; PF1959, Phosphoglycerate mutase; PF0215, Enolase; PF1188, Pyruvate kinase; PF0043, Phospho*enol*pyruvate synthetase; PF0074, short chain alcohol dehydrogenase; PF0073, ß-glucosidase; PF0594, Ornithine carbamoyltransferase; PF1602, Glutamate dehydrogenase; PF0825, Prolyl endopeptidase; PF1719, Protease I; PF0212, DNA polymerase; PF0287, Pyrolysin; PF1497, Alanine aminotransferase; PF0495, Reverse gyrase; PF1601, Leucine-responsive regulatory-like protein ; PF1532, NADH oxidase; PF0121, Aromatic aminotransferase; PF1253, Aromatic aminotransferase; PF1702, Aspartate aminotransferase is co-transcribed with putative chorismate mutase (PF1701); PF1790, Heat shock response regulator; PF1883, Small heat shock protein; PF1882, AAA+-ATPpase; PF1938, Maltodextrin binding protein

Table 4.2 Consensus sequences of archaeal promoter elements.

Archaeal groups	**TATA-box**[a]	**Transcription Factor B-responsive element**[a]	**Reference**
Halophiles	-29(T-T-T-W-W-W)-24	-[b]	(Soppa, 1999)
Methanogens	-30(Y-T-T-A-T-A-T-A)-23	-[b]	(Soppa, 1999)
Sulfolobus	-30(Y-T-T-T-T-A-A-A)-23	-36(R-N-W-A-A-W)-31	(Bell *et al.*, 1999; Soppa, 1999)
Pyrococcus	-29(T-T-W-W-W-A-W)-23	-36(V-R-A-A-A)-32	This study

IUPAC-code is used for ambiguous nucleotides (Cornish-Bowden, 1985).

[a] The numbers indicate the position of the nucleotides upstream the transcription start site.

[b] No consensus described.

genes appear to be transcribed as monocistronic messages (Siebers *et al.*, 2001; van der Oost *et al.*, 1998; Verhees *et al.*, 2001).

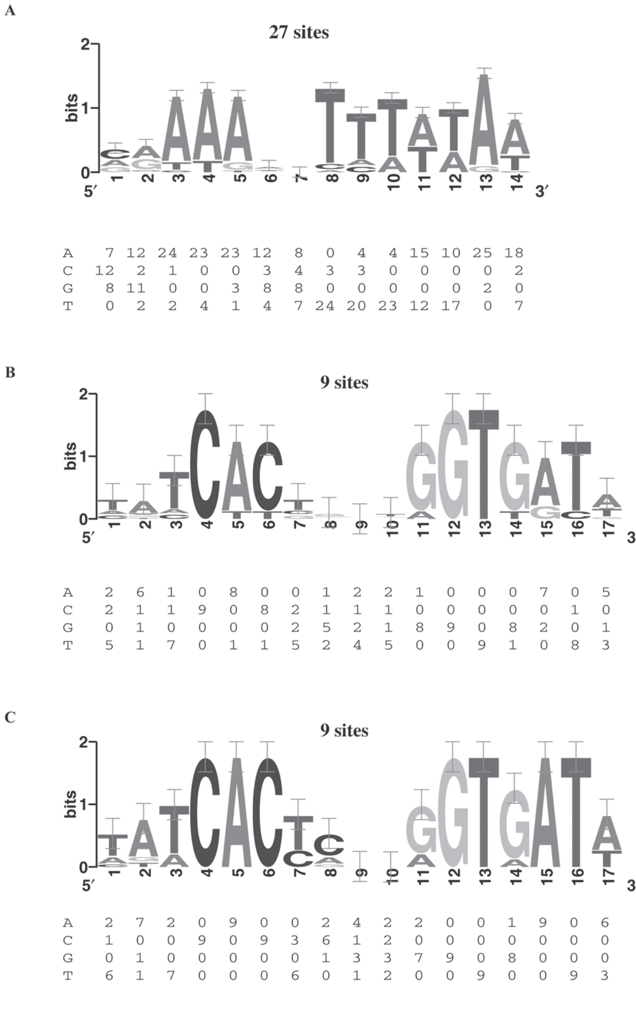
Transcription initiation sites of the glycolytic genes of *P. furiosus* were determined by primer extension analysis (Fig. 4.1). We were unable to determine the transcription initiation sites of the enolase and phosphophoglycerate mutase genes, maybe due to relative instability of these transcripts. Most of the investigated transcription start sites were found to be located at the first position, or immediately upstream of, a putative ribosome binding site (**GGTGAT**; the complementary 3'-end 16S rRNA sequence from *P. furiosus* is CGGCUCG**AUCACC**UCCU-3') (Fig. 4.1).

Gibbs Recursive Sampler algorithm was applied, on 27 sequences, to identify a pyrococcal Transcription Factor B-responsive element (BRE) and a TATA-box, with a maximum *a posteriori* (MAP) value of 100.9. The two conserved sequences most likely correspond to the BRE and the TATA-box (Fig. 4.2A) at the positions around –33/-34 and -26/-27 bases, respectively (Bell *et al.*, 1999) (Fig 4.1). Indeed, the archaeal TATA-Binding Protein (TBP) is known to bind to the TATA-box, generally centred at position -26/-27 bases. However, there can be some flexibility of 1 or 2 nucleotides in the spacing between the TATA-box and the transcription start site (Soppa, 1999). A consensus for the TATA-box sequences has been proposed for several archaeal groups (Table 4.2). Based on the comparison of the investigated *P. furiosus* promoter regions, the following TATA-box consensus is proposed: TTWWWAW (-29/-23) (W=T/A; Table 4.2). This consensus strongly resembles the consensus reported for halophiles. It is likely that the *Pyrococcus* TBP recognizes this sequence, based on the *in vitro* transcription studies of the glutamate dehydrogenase (Hethke *et al.*, 1996) and glyceraldehyde ferredoxin oxidoreductase genes (van der Oost *et al.*, 1998). The archaeal BRE plays a key role in directing the oriented assembly of the archaeal pre-initiation complex upon binding of Transcription Factor B (Bell *et al.*, 1998). A consensus sequence has been suggested for the 6-nucleotide BRE immediately upstream of the TATA-box for *Sulfolobus* (Table 4.2). Based on the analysis presented in this study, we propose a BRE consensus for *P. furiosus* of VRAAA (-36/-32) (V=C/G/A, R=G/A; Table 4.2), and for the overall BRE/TATA-box of VRAAA-N_2-

TTWWWAW (-36/-23).

Identification of a Thermococcales-Glycolytic-Motif in promoters of *P. furiosus* and *T. kodakaraensis*.

Detailed analysis of the aligned glycolytic promoter sequences revealed a motif that appears to be conserved in the glycolytic promoter sequences. All nine glycolytic promoters contain this motif (Fig. 4.1), with the assumption that phospho*enol*pyruvate synthetase instead of pyruvate kinase operates in the glycolytic direction in *Thermococcales* (see discussion below). The inverted repeat termed Thermococcales-Glycolytic-Motif (TGM) was detected in *P. furiosus* (MAP-value of 30.0), as well as in *T. kodakaraensis* (MAP-value of 55.9). The TGM consists of a conserved inverted repeat inter-spaced by five nucleotides (Fig. 4.2BC), with the consensus TATCAC-N_5-GTGATA. This putative *cis*-acting element could be involved in repression of the glycolytic genes at the transcriptional level, since it is in all cases located downstream of



A	7	12	24	23	23	12	8	0	4	4	15	10	25	18
C	12	2	1	0	0	3	4	3	3	0	0	0	0	2
G	8	11	0	0	3	8	8	0	0	0	0	0	2	0
T	0	2	2	4	1	4	7	24	20	23	12	17	0	7

A	2	6	1	0	8	0	0	1	2	2	1	0	0	0	7	0	5
C	2	1	1	9	0	8	2	1	1	1	0	0	0	0	0	1	0
G	0	1	0	0	0	0	2	5	2	1	8	9	0	8	2	0	1
T	5	1	7	0	1	1	5	2	4	5	0	0	9	1	0	8	3

A	2	7	2	0	9	0	0	2	4	2	2	0	0	1	9	0	6
C	1	0	0	9	0	9	3	6	1	2	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	3	3	7	9	0	8	0	0	0
T	6	1	7	0	0	0	6	0	1	2	0	0	9	0	0	9	3

Figure 4.2 (for color figure see Appendix I) Sequence logos and position-frequency matrices of (**A**) Transcription Factor B-responsive element and TATA-box, based on 27 *P. furiosus* promoter sequences and the Thermococcales-Glycolytic-Motif in promoter sequences, based on 9 glycolytic enzymes in *P. furiosus* (**B**) and *T. kodakaraensis* (**C**). The sequence logos were generated using WebLogo (Crooks *et al.*, 2004).

the predicted TATA-boxes (Fig. 4.1). It has indeed been shown that the transcription of the genes that encode GAPOR, phosphoglucose isomerase, fructose-1,6-bisphosphate aldolase and phospho*enol*pyruvate synthetase is higher during saccharolytic growth than under peptidolytic growth (Robinson *et al.*, 1994; Siebers *et al.*, 2001; van der Oost *et al.*, 1998; Verhees *et al.*, 2001). Furthermore, growth on tryptone inhibits glycolysis in the closely related organism *Thermococcus zilligii*, even after addition of glucose (Xavier *et al.*, 2000).

The intergenic regions of the complete genomes of *P. furiosus* and *T. kodakaraensis* were scanned with the TGM-position-frequency matrix and this resulted in 17 and 29 positive hits, respectively (Supplementary Table S4.1). Not only the TGM, but also its location is conserved in these two species (Supplementary Table S4.1). A major difference concerns the location of the TGM in the promoter sequences of the phosphoglucose isomerase in the two species (PF0196/17 bases in *Pyrococcus furiosus* vs. TK1111/67 bases upstream of the translation start in *Thermococcus kodakaraensis*). Further analysis of the position of the TGM

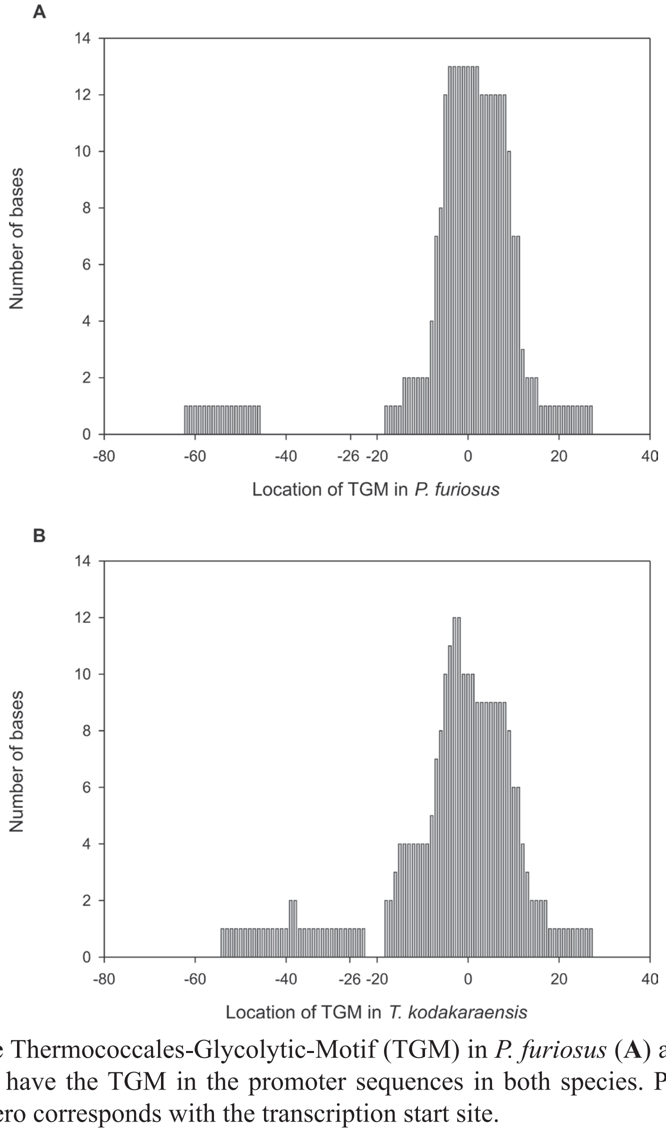


Figure 4.3 Location of the Thermococcales-Glycolytic-Motif (TGM) in *P. furiosus* (**A**) and *T. kodakaraensis* (**B**) of orthologous genes that have the TGM in the promoter sequences in both species. Putative TATA-boxes are centred at -26 bases and zero corresponds with the transcription start site.

in *P. furiosus* showed that it is mostly located downstream of the TATA-box and overlapped the transcription start site (Fig. 4.3), suggesting that it may be involved in the negative control of gene expression. We will only discuss the orthologous genes that have the TGM in the promoter sequences in both species.

Strikingly, the TGM was found in the promoter of the phospho*enol*pyruvate synthetase (*pps*) gene but not in the promoter of the pyruvate kinase (*pyk*) gene, which is of specific interest since it has been suggested that phospho*enol*pyruvate synthetase, rather than pyruvate kinase, might be operating in glycolytic direction in this archaeon (Sakuraba *et al.*, 1999). The observation that the promoter of the *pps* gene and other glycolytic genes contain the TGM corresponds with the induction of *pps* by maltose (Robinson *et al.*, 1994).

Several of the proteins encoded by genes with a TGM containing promoter are involved in starch and glucose metabolism. The presence of the motif in the promoter sequences of the (potential) operons that encode the ABC-transporter and a hydrolytic enzyme (amylase/amylopullulanase) involved in maltodextrin catabolism (PF1938-1933, TK1771-1775), may indicate that this putative *cis*-regulatory element is involved in transcriptional regulation of the whole cluster. In the promoter sequence of PF1938 of the maltodextrin operon, a recognition site of a transcriptional regulator (TrmB, PF1743) has recently been identified by footprint analysis (Lee *et al.*, 2005). Interestingly, the predicted TrmB binding site overlaps with the 3'-end of the TGM (Fig. 4.1). However, no TGM is present in the promoter area of the other target of TrmB in *P. furiosus*: the maltose/trehalose operon (PF1739-1744; encoding a specific ABC-transporter and the *trmB* gene). In *P. furiosus* three TrmB paralogs are present, whereas *T. kodakaraensis* only has a single homolog which is not orthologous with the characterized pyrococcal TrmB. The *T. kodakaraensis* TrmB-homolog (and its uncharacterized ortholog in *P. furiosus*) may be the regulator of the whole regulon (see discussion below).

Another interesting gene with a TGM in its promoter sequence is a phospho-sugar mutase (PF0588, TK1108). An experimental analysis of TK1108 revealed dual-specificity; catalyzing the isomerization of mannose-1-phosphate ↔ mannose-6-phosphate, as well as glucose-1-phosphate ↔ glucose-6-phosphate. The phosphoglucomutase activity and transcription of TK1108 was found to be higher in cells grown on starch vs. pyruvate and therefore it might be involved in starch degradation or intracellular glycogen synthesis (Rashid *et al.*, 2004).

Four α-glucan degrading enzymes genes also have the TGM in their promoter sequences in between BRE and TATA-box and the translation start (Supplementary Table S4.1). In fact, the gene expression of PF0272, PF0478 and PF0132 is up-regulated on maltose (26.0, 1.9 and 1.6-fold change in expression, respectively). According to (Lee *et al.*, 2006) all three enzymes are involved in starch and maltose metabolism, although PF0478 and PF0312 are not essential PF0477 on the other hand, is significantly down-regulated on maltose (5.7-fold) (Schut *et al.*, 2003). It is puzzling why this extracellular enzyme is down-regulated considering its annotated function and the good correlation between the TGM-positions. Schut *et al.* suggested that this enzyme is present during peptide fermentation, in case α-glucans become available; this extracellular amylase may degrade the polymers to dextrin-oligomers that are taken up and probably induce the glycolytic regulon. However, the presence of the TGM might suggest a different role.

The fructose-1,6-bisphosphatase (*fbp*) gene (PF0613, TK2164) is involved in gluconeogenesis and is down-regulated in cells grown on maltose (Schut *et al.*, 2003) and starch vs. pyruvate (Sato *et al.*, 2004). Interestingly, the TGM is found upstream of the BRE/TATA-box of the *fbp* gene, indicating that it might be an enhancer. This type of promoter architecture would resemble that of the two described archaeal transcription activators: *M. jannaschii* Ptr2 (Ouhammouch *et al.*, 2003) and *S. solfataricus* LysM (Brinkman *et al.*, 2002).

In contrast, the TGM is not present in promoter sequences of orthologous genes in *P. horikoshii* and *P. abyssi*. A comprehensive scan of all prokaryotic nucleotides with the TGM-matrix of *T. kodakaraensis* demonstrated the presence of the TGM in several other *Thermococcales* species (Supplementary Table S4.2) and the absence in other prokaryotes. The most likely evolutionary scenario would be the development of the TGM-related regulatory system in a common ancestor of *Pyrococcus* and *Thermococcus* and, after divergence of *P. furiosus*, its subsequent loss in the ancestor of *P. abyssi* and *P. horikoshii*. This corresponds with the relatively less complex catabolic capacity of the latter two species (Ettema *et al.*, 2001; Fukui *et al.*, 2005) and therefore it may not be necessary to regulate the carbohydrate metabolism, at the transcriptional level, as strict as in *P. furiosus* and in *T. kodakaraensis*.

Comparison of the predicted proteomes of the four completely sequenced species of the order *Thermococcales* revealed four putative regulators that are present in *P. furiosus* and *T. kodakaraensis* and absent in *P. abyssi* and *P. horikoshii*: (1) A sugar fermentation stimulation protein (SfsA; PF1198, TK0779), (2) a transcriptional regulator of the Lrp/AsnC family (PF0739, TK0834), (3) a transcriptional regulator of the PadR family, regulating phenolic acid decarboxylase in bacteria (PF1476, TK1494), and remarkably (4) a paralog (with a pairwise sequence identity of 26%) of the aforementioned TrmB regulator (PF0124, TK1769). SfsA has recently been suggested to be a nuclease rather than a regulator (Kosinski *et al.*, 2005), and the ligand specificity of members of the Lrp and PadR families do not suggest a link with sugar metabolism. The TrmB-like regulator, however, resembles the characterized regulator of trehalose/maltose metabolism in *Thermococcus litoralis* and *P. furiosus* (PF1743) (Lee *et al.*, 2003). An ortholog of the latter is absent in *T. kodakaraensis*, but the gene encoding a TrmB-like protein (TK1769) is clustered with the operon that encodes a maltodextrin ABC-transporter (TK1771-1775). Moreover, a TGM is present in the promoter sequence of TK1769 (Supplementary Table S4.1). We assume that the regulator is involved in recognizing the TGM, and subsequently modulates the expression of saccharolytic enzymes in *Thermococcales*. Experiments are ongoing to verify this hypothesis.

With the identification of the BRE and the TATA-box architecture in *P. furiosus,* the TGM and its conserved location and distribution in *P. furiosus* and *T. kodakaraensis,* it is concluded that the genes encoding proteins for glycolysis, sugar transport and α-linked sugar metabolism are part of the same regulon. This regulon is the largest that has yet been described for Archaea.

Supplementary material

For Table S4.1 and Table S4.2 see Appendix I.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Aravind, L., and Koonin, E. V. (1999). DNA-binding proteins and evolution of transcription regulation in the archaea. Nucleic Acids Res *27*, 4658-4670.

Baker, H. V. (1991). GCR1 of *Saccharomyces cerevisiae* encodes a DNA binding protein whose binding is abolished by mutations in the CTTCC sequence motif. Proc Natl Acad Sci U S A *88*, 9443-9447.

Bell, S. D., Jaxel, C., Nadal, M., Kosa, P. F., and Jackson, S. P. (1998). Temperature, template topology, and factor requirements of archaeal transcription. Proc Natl Acad Sci U S A *95*, 15218-15222.

Bell, S. D., Kosa, P. L., Sigler, P. B., and Jackson, S. P. (1999). Orientation of the transcription preinitiation complex in archaea. Proc Natl Acad Sci U S A *96*, 13662-13667.

Bell, S. D., Magill, C. P., and Jackson, S. P. (2001). Basal and regulated transcription in Archaea. Biochem Soc Trans *29*, 392-395.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Wheeler, D. L. (2005). GenBank. Nucleic Acids Res *33*, D34-38.

Borges, K. M., Bergerat, A., Bogert, A. M., DiRuggiero, J., Forterre, P., and Robb, F. T. (1997). Characterization of the reverse gyrase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol *179*, 1721-1726.

Brinkman, A. B., Bell, S. D., Lebbink, R. J., de Vos, W. M., and van der Oost, J. (2002). The *Sulfolobus solfataricus* Lrp-like protein LysM regulates lysine biosynthesis in response to lysine availability. J Biol Chem *277*, 29537-29549.

Brinkman, A. B., Dahlke, I., Tuininga, J. E., Lammers, T., Dumay, V., de Heus, E., Lebbink, J. H., Thomm, M., de Vos, W. M., and van Der Oost, J. (2000). An Lrp-like transcriptional regulator from the archaeon *Pyrococcus furiosus* is negatively autoregulated. J Biol Chem *275*, 38160-38169.

Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. Nucleic Acids Res *13*, 3021-3030.

Crooks, G. E., Hon, G., Chandonia, J. M., and Brenner, S. E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.

Diruggiero, J., Dunn, D., Maeder, D. L., Holley-Shanks, R., Chatard, J., Horlacher, R., Robb, F. T., Boos, W., and Weiss, R. B. (2000). Evidence of recent lateral gene transfer among hyperthermophilic archaea. Mol Microbiol *38*, 684-693.

Eggen, R. I., Geerling, A. C., Waldkotter, K., Antranikian, G., and de Vos, W. M. (1993). The glutamate dehydrogenase-encoding gene of the hyperthermophilic archaeon *Pyrococcus furiosus*: sequence, transcription and analysis of the deduced amino acid sequence. Gene *132*, 143-148.

Ettema, T., van der Oost, J., and Huynen, M. (2001). Modularity in the gain and loss of genes: applications for function prediction. Trends Genet *17*, 485-487.

Fukui, T., Atomi, H., Kanai, T., Matsumi, R., Fujiwara, S., and Imanaka, T. (2005). Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. Genome Res *15*, 352-363.

Geiduschek, E. P., and Ouhammouch, M. (2005). Archaeal transcription and its regulators. Mol Microbiol *56*, 1397-1407.

Halio, S. B., Blumentals, II, Short, S. A., Merrill, B. M., and Kelly, R. M. (1996). Sequence, expression in *Escherichia coli*, and analysis of the gene encoding a novel intracellular protease (PfpI) from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol *178*, 2605-2612.

Hethke, C., Geerling, A. C., Hausner, W., de Vos, W. M., and Thomm, M. (1996). A cell-free transcription system for the hyperthermophilic archaeon *Pyrococcus furiosus*. Nucleic Acids Res *24*, 2369-2376.

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

Kosinski, J., Feder, M., and Bujnicki, J. M. (2005). The PD-(D/E)XK superfamily revisited: identification of new members among proteins involved in DNA metabolism and functional predictions for domains of (hitherto) unknown function. BMC Bioinformatics *6*, 172.

Lee, H. S., Shockley, K. R., Schut, G. J., Conners, S. B., Montero, C. I., Johnson, M. R., Chou, C. J., Bridger, S. L., Wigner, N., Brehm, S. D*., et al.* (2006). Transcriptional and biochemical analysis of starch metabolism in the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol *188*, 2115-2125.

Lee, S. J., Engelmann, A., Horlacher, R., Qu, Q., Vierke, G., Hebbeln, C., Thomm, M., and Boos, W. (2003). TrmB, a sugar-specific transcriptional regulator of the trehalose/maltose ABC transporter from the hyperthermophilic archaeon *Thermococcus litoralis*. J Biol Chem *278*, 983-990.

Lee, S. J., Moulakakis, C., Koning, S. M., Hausner, W., Thomm, M., and Boos, W. (2005). TrmB, a sugar sensing regulator of ABC transporter genes in *Pyrococcus furiosus* exhibits dual promoter specificity and is controlled by different inducers. Mol Microbiol *57*, 1797-1807.

Lenhard, B., and Wasserman, W. W. (2002). TFBS: Computational framework for transcription factor binding site analysis. Bioinformatics *18*, 1135-1136.

Luesink, E. J., van Herpen, R. E., Grossiord, B. P., Kuipers, O. P., and de Vos, W. M. (1998). Transcriptional activation of the glycolytic *las* operon and catabolite repression of the *gal* operon in *Lactococcus lactis* are mediated by the catabolite control protein CcpA. Mol Microbiol *30*, 789-798.

Ouhammouch, M., Dewhurst, R. E., Hausner, W., Thomm, M., and Geiduschek, E. P. (2003). Activation of archaeal transcription by recruitment of the TATA-binding protein. Proc Natl Acad Sci U S A *100*, 5097-5102.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *33*, D501-504.

Ramseier, T. M., Bledig, S., Michotey, V., Feghali, R., and Saier, M. H., Jr. (1995). The global regulatory protein FruR modulates the direction of carbon flow in *Escherichia coli*. Mol Microbiol *16*, 1157-1169.

Rashid, N., Kanai, T., Atomi, H., and Imanaka, T. (2004). Among multiple phosphomannomutase gene orthologues, only one gene encodes a protein with phosphoglucomutase and phosphomannomutase activities in *Thermococcus kodakaraensis*. J Bacteriol *186*, 6070-6076.

Robinson, K. A., Bartley, D. A., Robb, F. T., and Schreier, H. J. (1995). A gene from the hyperthermophile *Pyrococcus furiosus* whose deduced product is homologous to members of the prolyl oligopeptidase family of proteases. Gene *152*, 103-106.

Robinson, K. A., Robb, F. T., and Schreier, H. J. (1994). Isolation of maltose-regulated genes from the hyperthermophilic archaeum, *Pyrococcus furiosus*, by subtractive hybridization. Gene *148*, 137-141.

Roovers, M., Hethke, C., Legrain, C., Thomm, M., and Glansdorff, N. (1997). Isolation of the gene encoding *Pyrococcus furiosus* ornithine carbamoyltransferase and study of its expression profile in vivo and in vitro. Eur J Biochem *247*, 1038-1045.

Sakuraba, H., Utsumi, E., Kujo, C., and Ohshima, T. (1999). An AMP-dependent (ATP-forming) kinase in the hyperthermophilic archaeon *Pyrococcus furiosus*: characterization and novel physiological role. Arch Biochem Biophys *364*, 125-128.

Sato, T., Imanaka, H., Rashid, N., Fukui, T., Atomi, H., and Imanaka, T. (2004). Genetic evidence identifying the true gluconeogenic fructose-1,6-bisphosphatase in *Thermococcus kodakaraensis* and other hyperthermophiles. J Bacteriol *186*, 5799-5807.

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

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

Siebers, B., Brinkmann, H., Dorr, C., Tjaden, B., Lilie, H., van der Oost, J., and Verhees, C. H. (2001). Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase. J Biol

Chem *276*, 28710-28718.

Soppa, J. (1999). Normalized nucleotide frequencies allow the definition of archaeal promoter elements for different archaeal groups and reveal base-specific TFB contacts upstream of the TATA box. Mol Microbiol *31*, 1589-1592.

Thompson, W., Rouchka, E. C., and Lawrence, C. E. (2003). Gibbs Recursive Sampler: finding transcription factor binding sites. Nucleic Acids Res *31*, 3580-3585.

Tuininga, J. E. (2004) Enzymology and bioenergetics of the glycolytic pathway of *Pyrococcus furiosus*., Wageningen University, Wageningen.

Tuininga, J. E., Verhees, C. H., van der Oost, J., Kengen, S. W., Stams, A. J., and de Vos, W. M. (1999). Molecular and biochemical characterization of the ADP-dependent phosphofructokinase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Biol Chem *274*, 21023-21028.

Uemori, T., Ishino, Y., Toh, H., Asada, K., and Kato, I. (1993). Organization and nucleotide sequence of the DNA polymerase gene from the archaeon *Pyrococcus furiosus*. Nucleic Acids Res *21*, 259-265.

van den Bogaard, P. T., Kleerebezem, M., Kuipers, O. P., and de Vos, W. M. (2000). Control of lactose transport, β-galactosidase activity, and glycolysis by CcpA in *Streptococcus thermophilus*: evidence for carbon catabolite repression by a non-phosphoenolpyruvate-dependent phosphotransferase system sugar. J Bacteriol *182*, 5982-5989.

van der Oost, J., Schut, G., Kengen, S. W., Hagen, W. R., Thomm, M., and de Vos, W. M. (1998). The ferredoxin-dependent conversion of glyceraldehyde-3-phosphate in the hyperthermophilic archaeon *Pyrococcus furiosus* represents a novel site of glycolytic regulation. J Biol Chem *273*, 28149-28154.

Verhees, C. H., Huynen, M. A., Ward, D. E., Schiltz, E., de Vos, W. M., and van der Oost, J. (2001). The phosphoglucose isomerase from the hyperthermophilic archaeon *Pyrococcus furiosus* is a unique glycolytic enzyme that belongs to the cupin superfamily. J Biol Chem *276*, 40926-40932.

Verhees, C. H., Kengen, S. W., Tuininga, J. E., Schut, G. J., Adams, M. W., De Vos, W. M., and Van Der Oost, J. (2003). The unique features of glycolytic pathways in Archaea. Biochem J *375*, 231-246.

Verhees, C. H., Koot, D. G., Ettema, T. J., Dijkema, C., de Vos, W. M., and van der Oost, J. (2002). Biochemical adaptations of two sugar kinases from the hyperthermophilic archaeon *Pyrococcus furiosus*. Biochem J *366*, 121-127.

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

Voorhorst, W. G., Eggen, R. I., Luesink, E. J., and de Vos, W. M. (1995). Characterization of the celB gene coding for β-glucosidase from the hyperthermophilic archaeon *Pyrococcus furiosus* and its expression and site-directed mutation in *Escherichia coli*. J Bacteriol *177*, 7105-7111.

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

Ward, D. E., de Vos, W. M., and van der Oost, J. (2002). Molecular analysis of the role of two aromatic aminotransferases and a broad-specificity aspartate aminotransferase in the aromatic amino acid metabolism of *Pyrococcus furiosus*. Archaea *1*, 133-141.

Ward, D. E., Donnelly, C. J., Mullendore, M. E., van der Oost, J., de Vos, W. M., and Crane, E. J., 3rd (2001). The NADH oxidase from *Pyrococcus furiosus*. Implications for the protection of anaerobic hyperthermophiles against oxidative stress. Eur J Biochem *268*, 5816-5823.

Ward, D. E., Kengen, S. W., van Der Oost, J., and de Vos, W. M. (2000). Purification and characterization of the alanine aminotransferase from the hyperthermophilic Archaeon *pyrococcus furiosus* and its role in alanine production. J Bacteriol *182*, 2559-2566.

Weinberg, M. V., Schut, G. J., Brehm, S., Datta, S., and Adams, M. W. (2005). Cold shock of a hyperthermophilic archaeon: *Pyrococcus furiosus* exhibits multiple responses to a suboptimal growth temperature with a key role for membrane-bound glycoproteins. J Bacteriol *187*, 336-348.

Xavier, K. B., da Costa, M. S., and Santos, H. (2000). Demonstration of a novel glycolytic pathway in the hyperthermophilic archaeon *Thermococcus zilligii* by (13)C-labeling experiments and nuclear magnetic resonance analysis. J Bacteriol *182*, 4632-4636.

Chapter 5

Hydrogenomics of the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus*

van de Werken, H. J. G., Verhaart, M. R. A., L., V. A., Willquist, K. U., Lewis, D. L., Nichols, J. D., Goorissen, H. P., Mongodin, E. F., Nelson, K. E., van Niel, E. W. J., Stams, A. J. M., Ward, D. E., de Vos, W. M., van der Oost, J., Kelly, R. M., and Kengen, S. W. M.
submitted for publication.

Abstract

Caldicellulosiruptor saccharolyticus is an extremely thermophilic, Gram-positive anaerobe, which ferments cellulose-, hemicellulose- and pectin-containing biomass to acetate, CO_2 and hydrogen. Its broad substrate range, high hydrogen-producing capacity, and ability to co-utilize glucose and xylose make this bacterium an attractive candidate for microbial bioenergy production. Here, the complete genome sequence of *C. saccharolyticus*, consisting of a 2,970,275 base pair circular chromosome encoding 2679 predicted proteins, is described. The genome reveals an extensive polysaccharide hydrolyzing capacity for cellulose, hemicellulose, pectin and starch, coupled to a large number of ABC transporters for monomeric and oligomeric sugar uptake. Components of the Embden-Meyerhof and the non-oxidative pentose phosphate pathways are all present, however, no evidence exists for an Entner-Doudoroff pathway. Catabolic pathways for a range of sugars, including rhamnose, fucose, arabinose, glucuronate, fructose, and galactose, were identified. These pathways presumably enable two different hydrogenase clusters to form H_2 from NADH or reduced ferredoxin. Whole-genome transcriptome analysis revealed significant upregulation of the glycolytic pathway and an ABC-type sugar transporter during growth on glucose and xylose, indicating that *C. saccharolyticus* co-ferments these sugars unimpeded by glucose-based catabolite repression. The capacity to simultaneously process and utilize a range of carbohydrates associated with biomass feedstocks represents a highly desirable feature of this lignocellulose-utilizing, biofuel-producing bacterium.

Introduction

Microbial hydrogen production from biomass has been recognized as an important route for renewable energy (USDOE2002; EC2002). For biohydrogen production from plant polysaccharides, high temperature microorganisms are well-suited, as anaerobic fermentation is thermodynamically favored at elevated temperature (Stams, 1994). The extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus* DSM 8903, a fermentative anaerobe initially isolated from wood in the flow of a thermal spring in New Zealand, first received attention for its capacity to utilize cellulose at its optimal growth temperature of 70 ºC (Rainey *et al.*, 1994). Further work showed that *C. saccharolyticus*: (1) can utilize a wide range of plant materials including cellulose, hemicellulose, starch and pectin, (2) has a very high hydrogen yield (almost 4 H_2 per mol of glucose) (de Vrije *et al.*, 2007; Kadar *et al.*, 2004; van Niel *et al.*, 2002), and (3) can ferment C5 and C6 sugars simultaneously. These features led to the development of bioprocessing schemes based on *C. saccharolyticus*. For example, H_2 production is now being investigated using a two-step process in which H_2 and acetate from biomass hydrolyzates are generated in one bioreactor, with the acetate fed to a second bioreactor to be used by phototrophic organisms (*Rhodobacter* spp.) to produce additional H_2 at the expense of light (Claassen and de Vrije, 2006). To provide a basis to fully exploit the biohydrogen producing capacity of *C. saccharolyticus*, its complete genome was sequenced and analyzed in conjunction with transcriptome information for this bacterium grown on glucose and xylose. Insights arising from this effort reveal that *C. saccharolyticus* has the capacity to process and utilize a broad range of sugars, ultimately forming hydrogen from their catabolism.

Results

General features and comparative genomics of the genome of *C. saccharolyticus*

The genome of *Caldicellulosiruptor saccharolyticus* DSM 8903/ATCC 43494 consists of one circular chromosome of 2,970,275 base pairs (bp), which has a G+C-content of 35.3% (Table 5.1). The gene locations of the 2679 predicted coding sequences on the two strands reflect the correlation between the direction of transcription and replication, and show a chromosome

Table 5.1 General features of *Caldicellulosiruptor saccharolyticus* genome

Length chromosome (bp)	2,970,275
G+C content (%)	35.3
Coding density (%)	86.3
Total number of protein-coding genes	2679
Average length of the protein-coding genes (b)	958
Total number of pseudogenes	92
Total number of tRNA genes	46
Total number of rRNA genes	9 (3 operons)
CRISPR-loci	9

with two unequal replichores (Fig. 5.1). In addition, the GC-skew analysis confirms the huge size difference of both replication arms, which might be attributed to a recent major inversion event. Apart from protein-coding genes, which are classified according to the COG system (Table 5.2), the chromosome harbors three ribosomal RNA operons and 46 tRNA genes with 41 different anticodons. These anticodons encode for all the 20 canonical amino acids. Like in many prokaryotes, the chromosome of *C. saccharolyticus* contains Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR). CRISPR are DNA repeats separated by highly variable intervening sequences (spacers) and accompanied by CRISPR-associated (CAS) genes. The CRISPR and CAS proteins have been proposed to function as a defense mechanism against bacteriophages (Barrangou *et al.*, 2007). With nine CRISPR loci and three different CAS genes, *C. saccharolyticus* is well-equipped to fend off bacteriophages.

The complete genome sequence confirms the phylogenetic position of *C. saccharolyticus* as member of the class Clostridia and reveals *Thermoanaerobacter tengcongensis* (whose genome sequence has also been completed) as closest relative (Supplementary Table S5.1). The genome of *C. saccharolyticus* was compared to two thermophilic relatives: *Clostridium thermocellum* and *T. tengcongensis* (Bao *et al.*, 2002), as well as to distantly related hyperthermophiles, the

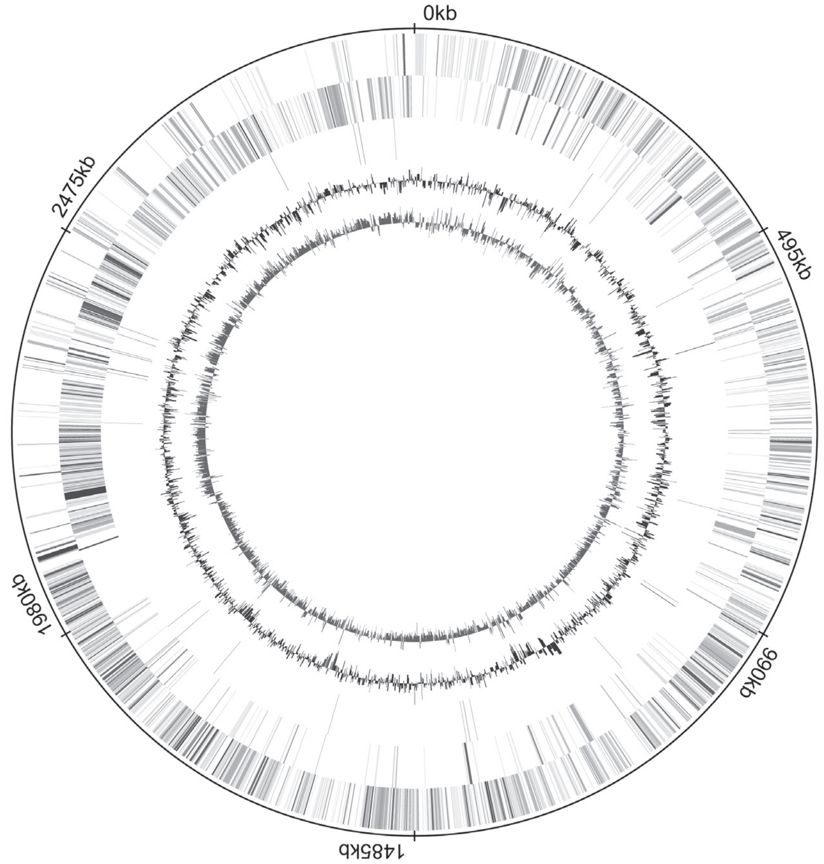


Figure 5.1 (for color figure see Appendix II) Circular representation of the Caldicellulosiruptor saccharolyticus chromosome. From the outer circle to the inner circle (1) genomic position in kilobases (kb) (2) coding sequences on the positive and (3) negative strand, which are colored according to the Clusters of Orthologous Groups of proteins (COG) functional categories, (4) tRNA genes (5) GC% (blue) (5) GC-skew (red). The Microbial Genome Viewer was used to make the circular chromosome wheel (Kerkhoven *et al.*, 2004).

Table 5.2 Functional categories of predicted open reading frames in the genomes of hydrogen producing organisms

	COG functional categories[a]	**Number of proteins in genomes**[b]				
		Caldicellulosiruptor saccharolyticus	*Clostridium thermocellum*	*Thermoanaerobacter tengcongensis*	*Thermotoga maritima*	*Pyrococcus furiosus*
	Information storage and processing					
B	Chromatin structure and dynamics	2	1	2	1	4
L	Replication, recombination and repair	222	252	149	89	109
A	RNA processing and modification	0	0	0	0	2
K	Transcription	134	174	141	82	80
J	Translation, ribosomal structure and biogenesis	147	165	150	135	166
	Cellular processes and signaling					
D	Cell cycle control, cell division, chromosome partitioning	35	38	40	21	18
N	Cell motility	71	95	67	58	13
M	Cell wall/membrane/envelope biogenesis	107	172	110	76	61
Z	Cytoskeleton	3	1	0	0	0
V	Defense mechanisms	48	40	42	27	29
U	Intracellular trafficking, secretion, and vesicular transport	42	60	46	39	20
O	Posttranslational modification, protein turnover, chaperones	59	89	81	55	55
T	Signal transduction mechanisms	125	170	122	73	19
	Metabolism					
E	Amino acid transport and metabolism	166	166	206	181	158
G	Carbohydrate transport and metabolism	213	144	160	166	93
H	Coenzyme transport and metabolism	101	102	67	62	89
C	Energy production and conversion	111	115	130	119	128
P	Inorganic ion transport and metabolism	73	91	96	113	95
I	Lipid transport and metabolism	34	46	55	31	25
F	Nucleotide transport and metabolism	56	62	62	54	52
Q	Secondary metabolites biosynthesis, transport and catabolism	14	18	27	15	11
	Poorly characterized					
S	Function unknown	177	187	173	137	189
R	General function prediction only	228	261	249	204	275
	Not in COG	655	985	615	305	322

[a]Gene classification according to Integrated Microbial Genomes (IMG) system (Markowitz *et al.*, 2006) using the functional classification of Clusters of Orthologous Groups of proteins (COG) (Tatusov *et al.*, 2003)
[b]Number of protein-coding genes in each category without pseudogenes.

bacterium *Thermotoga maritima* (Nelson *et al.*, 1999) and the archaeon *Pyrococcus furiosus* (Robb *et al.*, 2001). These microorganisms have small to moderate genome sizes and also produce hydrogen while growing on a range of carbohydrates (Supplementary Table S5.2). The genomic distribution of proteins into COG categories is comparable for this group of species (Table 5.2). However, one major difference is that both *C. thermocellum* and *C. saccharolyticus* have many more transposases and transposase derivatives than the other species, namely, 99 and 93, respectively. The genomes of *T. maritima*, *P. furiosus* and *T. tengcongensis*, harbor 11, 17 and 43 transposases and transposase derivatives, respectively. Therefore, *C. saccharolyticus* and *C. thermocellum* have a high number of proteins, 222 and 252, respectively, in the category 'Replication, recombination and repair' (L). Furthermore, the *C. saccharolyticus* genome harbors the largest number of carbohydrate transport and metabolism genes in this group. In fact, the *C. saccharolyticus* genome contains at least 177 ABC-transporter genes, outnumbering the 165 identified in *T. maritima* (Conners *et al.*, 2005; Nelson *et al.*, 1999). The *C. saccharolyticus* genome contains a reverse gyrase gene (Csac_1580), the product of which induces positive supercoiling of DNA (Kikuchi and Asai, 1984). Reverse gyrase is regarded as a molecular marker of hyperthermophilicity, and therefore distinguishes *C. saccharolyticus* from *C. thermocellum*, which lacks reverse gyrase. Despite the fact that *C. saccharolyticus* was described as a non-motile organism, a set of flagella structure, biogenesis and chemotaxis genes were detected; it is not clear whether these genes are functional, since one is interrupted by a stop codon (pseudogene Csac_1277). *C. saccharolyticus* has a nitrogen-fixation cluster (Csac_2461-2466) and many sporulation genes; neither of these phenotypic properties have been described for this bacterium.

Central carbon metabolism

C. saccharolyticus is able to metabolize a wide variety of carbohydrates, including the monosaccharides D-glucose, D-xylose, D-fructose, D-galactose, D-/L-arabinose, D-mannose, L-rhamnose and L-fucose, but also α- and β-linked di- and poly-saccharides, including maltose, starch, pullulan, sucrose, trehalose, amorphous and micro-crystalline cellulose, xylan, locust bean gum and pectin (Rainey *et al.*, 1994). Once hydrolyzed, sugars are channeled to the central catabolic pathways (Fig. 5.2). The genome sequence reveals components of a complete Embden-Meyerhof (EM) pathway, including a ROK family glucokinase (Csac_0778), 6-phosphofructokinase, (Csac_2366/1830), a bifunctional phosphoglucose/phosphomannose isomerase (Csac_1187), fructose-1,6-bisphosphate aldolase (Csac_1189), pyruvate kinase (Csac_1831) as well as pyruvate-phosphate dikinase (PPDK) (Csac_1955). Also a *gapA* operon is evident, consisting of glyceraldehyde-3-phosphate (GAP) dehydrogenase (Csac_1953), the fusion protein phosphoglycerate kinase/triose-phosphate isomerase (Csac_1952), phosphoglycerate mutase (Csac_1951), and enolase (Csac_1950) (Fig 5.2). However, the

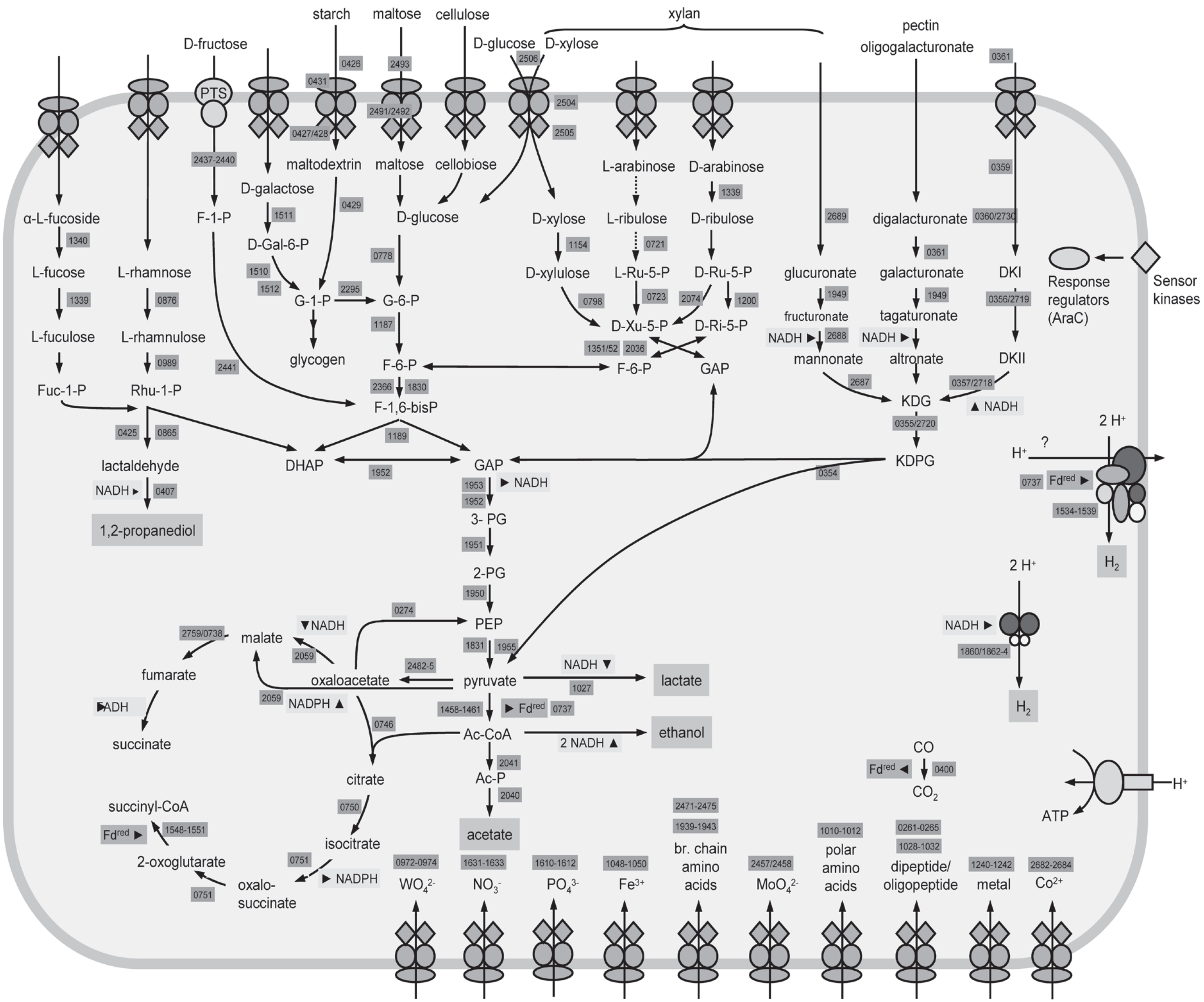


Figure 5.2 (for color figure see Appendix III) An overview of the carbon metabolism and transport systems in *Caldicellulosiruptor saccharolyticus*. The identity of the various ABC-type sugar transporters is not known. Secondary transport systems may be involved as well.

oxidative branch (ox) of the Pentose Phosphate Pathway (PPP) and the Entner-Doudoroff (ED) pathway were not detected, which is consistent with previous reports using ^{13}C-NMR (de Vrije *et al.*, 2007). The absence of the ox-PPP, however, raises questions about how NADPH is produced for biosynthesis. The only other obvious NADP-producing reaction is isocitrate dehydrogenase (Csac_0751). However, based on its sequence homology the isocitrate dehydrogenase is likely to produce NADH instead of NADPH. Also, no obvious homolog to an NADPH-producing glyceraldehyde-3-phosphate dehydrogenase can be identified, as has been reported for *Streptococcus* species and some clostridia (Boyd *et al.*, 1995). Furthermore, no ferredoxin:NADPH reductase homolog is present, although such activity has been measured in some *Thermoanaerobacter* spp. (Hyun *et al.*, 1985).

Xylose, a major constituent of hemicellulose, is funneled by a putative xylose isomerase (Csac_1154) and xylulokinase (Csac_0798) into the non-oxidative branch (nox) of the PPP. The nox-PPP uses ribulose-phosphate 3-epimerase (Csac_2074), ribose-5-phosphate isomerase (Csac_1200), the N-terminal (Csac_1351) and C-terminal transketolase (Csac_1352) and transaldolase (Csac_2036) to produce the EM intermediates fructose-6-phosphate and glyceraldehyde-3-phosphate. Galactose also enters the EM via the Leloir-pathway, which includes galactokinase (Csac_1511), galactose-1-phosphate uridylyltransferase (Csac_1510), UDP-glucose 4-epimerase (Csac_1512) and phosphoglucomutase (Csac_2295). Strikingly, none of the established types of fructose-bisphosphatase (Class I to IV; (Sato *et al.*, 2004)) are evident in the *C. saccharolyticus* genome. Since fructose-bisphosphatase is an essential enzyme of the gluconeogenesis, *C. saccharolyticus* presumably uses a novel phosphatase. Moreover, a gene for the gluconeogenic PEP synthetase is also missing, although the conversion of pyruvate to PEP could occur via the reversible PPDK (Csac_1955) or via oxaloacetate.

Pyruvate, the end product of the EM-pathway, is subsequently decarboxylated to acetyl-CoA by pyruvate:ferredoxin oxidoreductase (POR). *C. saccharolyticus* contains three 2-oxoacid:ferredoxin oxidoreductase enzyme complexes (Csac_2248-2249, 1458-1461 and Csac_1548-1551). According to transcriptional response information (*vide infra*), the true POR is probably encoded by Csac_1458-1461. Acetyl-CoA is used to generate acetate and ATP (Csac_2040/2041), or it enters the tricarboxylic acid (TCA) cycle for biosynthetic purposes. The TCA cycle in *C. saccharolyticus* is incomplete, with an oxidative branch to succinyl-CoA catalyzed by a citrate (Re)-synthase (Csac_0746), aconitate hydratase (Csac_0750), isocitrate dehydrogenase (Csac_0751) and the 2-oxoglutarate:ferredoxin oxidoreductase complex (1548-1551). In the reductive direction, only orthologs of the subunits of fumarate hydratase were detected with a high level of confidence (Csac_2759/Csac_0738). Malate dehydrogenase (oxaloacetate-decarboxylating) (Csac_2059) may be used to generate malate directly from pyruvate instead from oxaloacetate. Fumarate reductase, however, could not be identified, which is in agreement with the lack of this enzyme in related clostridia. Besides the malate dehydrogenase, TCA metabolites could be replenished by a putative sodium pump oxaloacetate decarboxylase enzyme complex (Csac_2482-2485).

Polysaccharide degrading enzymes

The capacity of *C. saccharolyticus* to hydrolyze a broad range of polysaccharides prior to fermentation differentiates this bacterium from many thermophilic anaerobes. Indeed, the genome of *C. saccharolyticus* encodes a wide range of carbohydrate active enzymes (Supplementary Table S5.3). These carbohydrate-utilizing enzymes are often clustered on the chromosome and can be assigned to substrate specific catabolic pathways for cellulose, hemicellulose and, to a lesser extent, starch and pectin. The α-1,4-glucan polymers, for instance, can be transported into the cell using the maltodextrin ABC-transport system proteins (Csac_0427-0428/0431). An intracellular α (Csac_0426) and a 1,4-α-glucan phosphorylase (Csac_0429) further degrade the intracellular maltodextrins, releasing glucose-1-phosphate. Remarkably, a transcriptional regulator of the LacI family (Csac_0430) is also in this maltodextrin cluster and is, therefore, a good candidate for controlling expression of this maltodextrin-degrading pathway at the transcriptional level. In addition, a GCAAACGTTTGC consensus sequence was found in upstream sequences of this transport cluster and several starch-degrading enzymes, such as an α-amylase precursor (Csac_0408), an oligo-1,6-glucosidase (Csac_2428), a pullulanase (Csac_0689), a 4-α-glucanotransferase (Csac_0203), and a putative glucan 1,4-α-glucosidase (Csac_0130). The consensus sequence resembles the binding site (CGCAAACGTTTGCGT) of the maltose/maltodextrin transcriptional repressor MalR from the Gram-positive *Streptococcus pneumoniae* (Nieto *et al.*, 1997). Besides this putative starch-degrading regulon, *C. saccharolyticus* has a glycogen metabolic cluster (Csac_0780-0784), a maltose ABC-transport system (Csac_2491-3), and a second pullulanase (Csac_0671). Taken together, *C. saccharolyticus* is well-equipped for starch utilization.

An important feature of *C. saccharolyticus* is its ability to produce H_2 not only from α-linked polymers, but also from complex β-linked glycans, such as cellulose, hemicellulose, laminarin and galactomannan. Growth on cellulosic substrates is rare among (hyper)thermophilic microorganisms. *C. saccharolyticus* does not metabolize cellulose by means of a cellulosome (Te'o *et al.*, 1995). For example, typical molecular components of a cellulosome, *i.e.*, dockerin domains and scaffolding proteins, were not identified in the genome. Nevertheless, a gene cluster (Csac_1076-1081) containing cellulase precursors is present. These highly modular cellulases are potentially capable of degrading this plant polysaccharide (Bergquist *et al.*, 1999) (Supplementary Table S5.3). Moreover, another gene cluster (Csac_1089-1091) and an extracellular cellulase (Csac_0678) may assist in completely hydrolyzing cellulose to glucose.

C. saccharolyticus has an enzyme system to cleave the glycoside bonds and hydrolyze ester bonds in hemicellulose (Csac_2404-2411). These mostly extracellular enzymes, which are variable in domain composition as well (Bergquist *et al.*, 1999), might be co-expressed with a smaller putative xylan-utilizing cluster (Csac_0203-0205). This latter cluster was not significantly up-regulated on xylose, in contrast to genes in the former cluster. Furthermore, putative genes that encode enzymes to degrade galactomannan (Csac_0663-0664), galactoarabinan (Csac_1560-1562) and laminarin (Csac_2548) can be identified.

The plant cell wall component pectin consists of α-1,4-linked D-galacturonic acid backbone, sometimes interspersed by L-rhamnose, and side chains made of monosaccharides, such as D-galactose, D-xylose and L-arabinose (Ridley *et al.*, 2001). Degradation of the main pectin component, D-galacturonate, requires a galacturonate isomerase, a tagaturonate reductase, and an altronate dehydratase to form 2-keto-3-deoxygluconate (KDG). Galacturonate isomerization may occur by glucuronate isomerase (Csac_1949). However, tagaturonate reductase and altronate dehydratase were not detected in the genome of *C. saccharolyticus*. Apparently, novel enzymes or a novel pathway are responsible for the degradation of galacturonate. In contrast, a gene cluster for the conversion of glucuronic acid to KDG (Csac_2686-2689) can be identified, and includes fructuronate reductase, mannonate dehydratase, a putative β-galactosidase/β-glucuronidase, and an α-glucuronidase. Glucuronic acid is a common substituent of xylan. Enzymes for the subsequent conversion of KDG to pyruvate and GAP, *viz.* KDG kinase (Csac_0355 or Csac_2720) and KDG-6-phosphate aldolase (Casc_0354) are present as well. The encoding genes of these last two steps are clustered with genes (Csac_0356-0357 and Csac_2718-2719) that both metabolize 5-keto-4-deoxyuronate (DK-I), an unsaturated cleavage product from pectate, to KDG. The enzymes that are able to hydrolyze the pectate backbone and the side chains (*e.g.*, unsaturated rhamnogalacturonyl hydrolase (Csac_0360), galacturan 1,4- α -galacturonidase (Csac_0361), β-galactosidase (Csac_0362) and a glycoside hydrolase with unknown substrate specificity (Csac_0363)) are in proximity to these KDG metabolic enzymes as well. However, neither a pectate lyase nor a methylesterase could be definitively identified in the genome; although Csac_2721/2728 might be candidates for a pectate lyase based on distant homology to known lyases.

C. saccharolyticus is also able to grow on L-rhamnose and on L-fucose, thereby producing 1,2-propanediol as end product (unpublished data). A putative rhamnose catabolic pathway can be assigned that generates dihydroxy-acetone phosphate and 1,2-propanediol catalyzed by a L-rhamnose isomerase (Csac_0876), a putative L-rhamnulokinase (Csac_0989), a L-rhamnulose-1-phosphate aldolase (Csac_0865) and a putative lactaldehyde reductase (Csac_0407). Fucose can be processed by a similar pathway, using the aforementioned lactaldehyde reductase, and yet to be identified versions of L-fuculokinase, a bifunctional L-fucose isomerase/D-arabinose isomerase (Csac_1339) and fuculose-1-phosphate aldolase (Csac_0425).

Fermentation products

Reducing equivalents are produced at the level of NAD and ferredoxin (Csac_0737). Since *C. saccharolyticus* can produce almost 4 H_2 per mol of glucose (de Vrije *et al.*), both NADH and reduced ferredoxin should ultimately be able to transfer their reducing equivalents to protons to form hydrogen. In the genome, two hydrogenase gene clusters could be identified, which are very similar to the two related clusters in *Thermoanaerobacter tengcongensis* (Soboh *et al.*, 2004). The first cluster (Csac_1534-1539) encodes subunits of a Ni-Fe hydrogenase (EchA-F) and various genes required for maturation of the hydrogenase complex (HypA-F; Csac_1540-1545). For *T. tengcongensis,* this Ni-Fe hydrogenase is ferredoxin-dependent, membrane-bound, and may act as a proton pump to generate a proton motive force. The second cluster (Csac_1860-1864) codes for a Fe-only hydrogenase (HydA-D), which is NAD-dependent and located in the cytoplasm, similar to the case for *T. tengcongensis* (Soboh *et al.*, 2004). Hydrogenases that form H_2 directly from NADH are unusual, and make an NAD:Fd oxidoreductase (Nfo) redundant. Nfo's (also known as Rnf) are membrane-bound multi-subunit complexes that use or create a Na^+-gradient coupled to the transfer of reducing equivalents between NADH and ferredoxin (Boiangiu *et al.*, 2005). An Nfo-cluster has been identified in the genomes of *C. thermocellum, T. maritima* and *T. ethanolicus*, but not in *T. tengcongensis* and *C. saccharolyticus*. The absence of an Nfo in *C. saccharolyticus* also implies that under elevated levels of H_2, reduced ferredoxin may either not be used to produce NADH or that a novel type of enzyme (complex) performs this reaction. Altogether, information available suggests that *C. saccharolyticus* is able to produce hydrogen from ferredoxin, but can also do this directly from NADH. Production of hydrogen would seem to be preferable, because under these conditions all pyruvate is converted to acetate (and CO_2), which is coupled to the synthesis of ATP.

When the hydrogen partial pressure (pH_2) becomes too high, hydrogen formation from NADH is no longer thermodynamically favorable. In that case, NADH is oxidized through the formation of lactate or ethanol. A gene for a lactate dehydrogenase can be identified (Csac_1027), but genes for acetaldehyde dehydrogenase and alcohol dehydrogenase were not obvious. In *T. tengcongensis* and *T. ethanolicus*, ethanol formation is NADPH-dependent and catalyzed by a bifunctional ADH acetyl-CoA thioesterase; this enzyme also has a homolog in *C. saccharolyticus* (Csac_0395).

A third small hydrogenase-like cluster could be detected in the *C. saccharolyticus* genome, composed of four genes encoding two NADH-binding proteins (Csac_0619-0620), a molybdopterin oxidoreductase containing NAD and 4Fe-4S binding regions (Csac_0621), and an iron-containing alcohol dehydrogenase (Csac_0622). The function of this cluster is yet unknown.

Transport systems

As mentioned earlier, there are a number of genes involved in ABC transporters found in the *C. saccharolyticus* genome including the previously noted carbohydrate specific maltodextrin ABC-transport system (Csac_0427-0428/0431) and the maltose ABC-transport system (Csac_2491-3). As is the case with certain *T. maritima* maltose transporters, both sets of these transport proteins lack ATP-binding subunits. For many bacteria, the intracellular ATPase used in the system is not encoded within the same operon. Both of these set of ABC transporters are located downstream from a two-component system (TCSs) of a sensor histidine protein kinase and a response regulator. In *C. saccharolyticus* ~50% of the ABC carbohydrate transport systems are located near TCSs on the chromosome.

Comparative analysis of *C. saccharolyticus* sugar binding proteins (SBPs) revealed that about 2/3 belong to COG1653. This category includes the CUT1 subfamily members TM0432, TM0595 and TM1855 that transport a variety of di- and oligosaccharides, such as maltose. More than half of COG1653 members are proximate in genomes to glycoside hydrolases, supporting their designation as ABC transporters involved in carbohydrate utilization. Putative

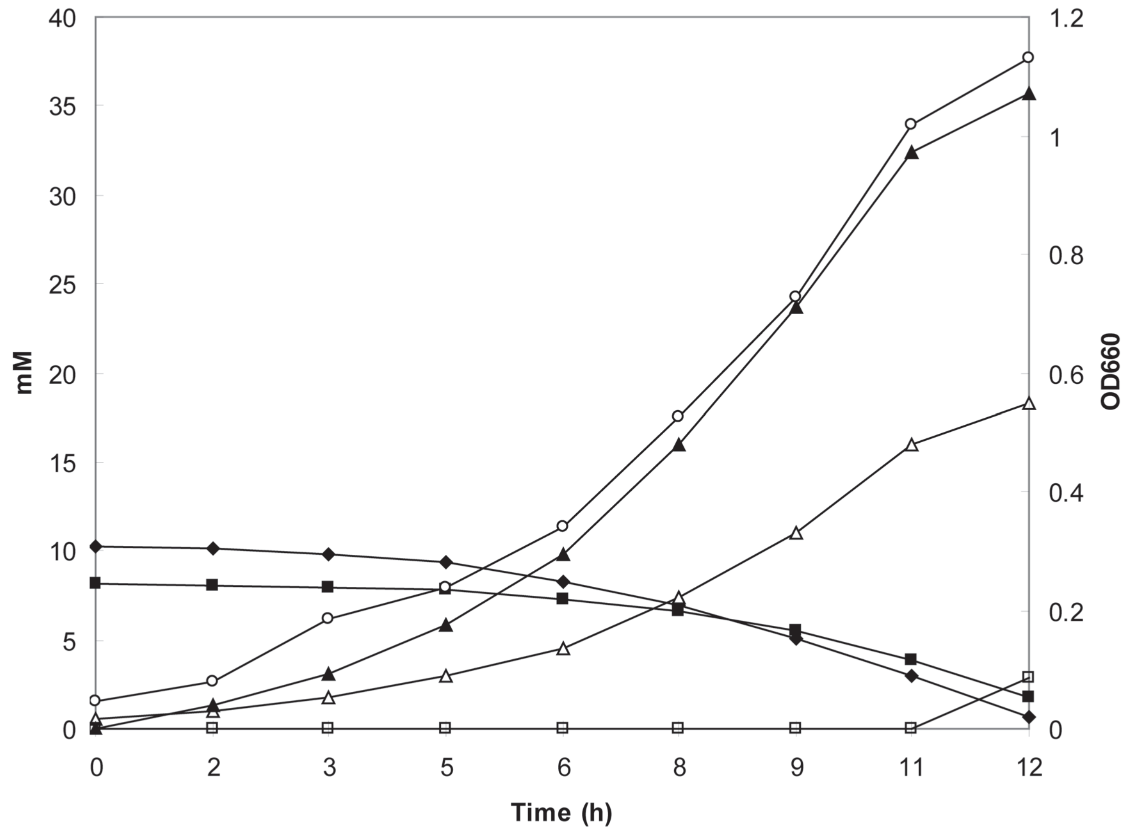


Figure 5.3 Growth of Caldicellulosiruptor saccharolyticus on a xylose:glucose mixture (1:1 (w/w)). ○ = optical density at 660 nm (OD660); ▲= hydrogen; Δ = acetate; ■ = glucose; ♦ = xylose; □ = lactate.

SBPs Csac_0242, Csac_0391, Csac_2326 and Csac_2507 belong to COG1879. Csac_2506 and Csac_2510 are associated with the xylose transport specific COG4213. As is the case with *T. maritima*, a few putative SBPs (Csac_0261 and Csac_4166) annotate as peptide transporters, although their actual function is unknown. Components of phosphotransferase systems (PTSs) have been identified in *C. saccharolyticus* (although only one set of carbohydrate-specific EII), along with a few putative members of the major facilitator superfamily (Csac_0685, Csac_0786, Csac_1100, Csac1170 and Csac_2298). However, it is likely that carbohydrate utilization proceeds mainly through ABC transporters.

Transcriptional regulation

The ability of *C. saccharolyticus* to utilize many different carbohydrates suggests a tight regulation among the pathways. Many carbohydrate utilization pathways in the genome appear to be regulated at the transcriptional level. Apart from the RNA polymerase core enzyme subunits (Csac_2259/2085/0951/0952), this Gram-positive species has 12 different σ-factors to construct the RNA polymerase holoenzyme. In addition, many of the sugar transcriptional regulators are present in multiple copies in the genome, *e.g.* nine proteins from the LacI family, six proteins from the DeoR family and eight from the GntR family, as well as 19 receiver proteins from a two-component system with a helix-turn-helix AraC domain. The latter are always clustered with sugar transporters and sugar hydrolytic enzymes.

Carbon catabolite repression (CCR) by glucose was not observed in *C. saccharolyticus* (Fig. 5.3). Nevertheless, some indicators of a Carbon Control Protein A (CcpA) dependent CCR in Gram-positives are present in the genome: (i) a histidine-containing phosphocarrier (HPr) (Csac_2438) that is in proximity to the only phospho*enol*pyruvate-dependent phosphotransferase system (PTS), which is fructose specific; (ii) A HPr(Ser) kinase (Csac_1186); (iii) a catabolite repression HPr (CrH) (Csac_1163) and nine members of the CcpA-containing LacI family. Binding sites for a putative CcpA, the catabolite-responsive element (*cre*), could not be identified. The *Bacillus subtilis* consensus sequence WWTGNAARCGNWWWCAWW (Miwa *et al.*, 2000) is, for instance, detected only twice: (1) In the upstream region of the aforementioned α-amylase precursor (Csac_0408) where it overlaps with the putative MalR binding site. (2) In the middle of the gene encoding the fumarate hydratase subunit α (Csac_2759). Nevertheless, in *C. saccharolyticus*, CCR is probably present, although the metabolite that induces this repression is unknown.

Besides global regulation through CCR, many local transcriptional regulators control the expression of carbohydrate metabolic pathways. Several orthologous transcriptional regulators were identified in *C. saccharolyticus*. The central glycolytic genes regulator (CggR) (Csac_1954), for instance, represses the transcription of the *gapA* operon (Ludwig *et al.*, 2001), while the FruR (Csac_2442) controls the fructose operon (Barriere *et al.*, 2005). Based on the fact that many transcriptional regulators are in proximity to their target operons, putative functions could be assigned to: an α-linked glucan transcriptional regulator (Csac_0430), a regulator of the oxidative-branch of the TCA cycle to oxoglutarate (Csac_0752), a repressor of the L-arabinose metabolism (Csac_0722), and a putative response regulator receiver protein of the glucuronate degradation (Csac_2690).

Transcriptome analysis Caldicellulosiruptor saccharolyticus

One of the beneficial features of *C. saccharolyticus* for hydrogen production is its ability to degrade cellulosic substrates as well as hemicellulose. Moreover, mixtures of glucose and xylose can be fermented simultaneously (Fig. 5.3) suggesting that classical CCR by glucose

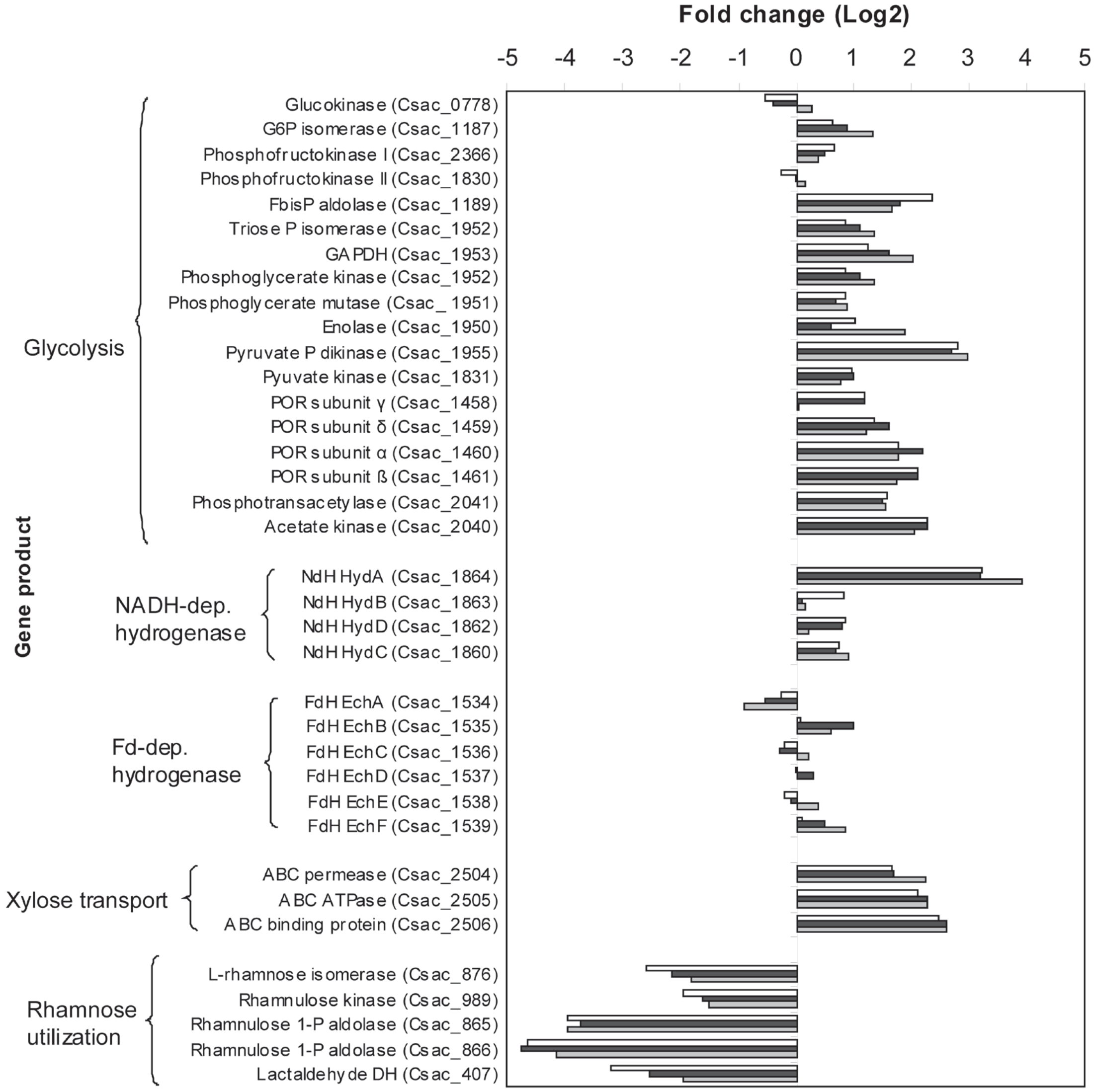


Figure 5.4 Intensity ratio of transcript levels of selected genes that responded to growth on glucose (white), xylose (gray) or a mixture of glucose and xylose (black), compared to rhamnose. Ratios are expressed as $\log_2$ value. GAPDH, glyceraldehyde-3-phosphate dehydrogenase; POR, pyruvate:ferredoxin oxidoreductase; NdH, NAD-dependent hydrogenase; FdH, ferredoxin-dependent hydrogenase.

does not occur. To elucidate the central carbon metabolic pathways and their regulation, transcriptome analysis was performed after growth on glucose, xylose and a 1:1 mixture of both substrates. L-Rhamnose, which was likely to follow another pathway, was used as a reference substrate. The transcriptional data clearly show that glucose, xylose and the glucose:xylose mixture all trigger up-regulation of genes in the EM pathway, when compared to rhamnose (Fig 5.4; Supplementary Table S5.4). In particular, the fructose-bisphosphate aldolase, GAP dehydrogenase, PPDK and POR are significantly stimulated. The ultimate acetate-forming acetate kinase is also highly up-regulated. A catabolic role for PPDK is intriguing, since it normally is associated with gluconeogenesis (as in propionic acid bacteria and plants), and PEP is usually converted by pyruvate kinase. However, homologs of PPDK are also present in related clostridia and *Thermoanaerobacter* species.

It is worth noting that growth on glucose, xylose or both sugars all trigger transcription of the gene encoding a xylose-specific ABC transport system (Csac_2504-2506) (Fig. 5.4), suggesting that glucose and xylose are transported by the same uptake system. Moreover, none of the identified putative CCR genes (*vide infra*) were differentially transcribed, confirming the fact that catabolite repression by glucose was not a factor.

Transcriptional response to growth on monosaccharides enabled the identification of genes and groups of adjacent genes (gene clusters), that were specifically up-regulated in response to either glucose or xylose. On glucose, several genes coding for α-glucan hydrolases responded. The most striking observation, however, was the up-regulation of an entire gene cluster (Csac_1991-2000) involved in purine synthesis, that was not observed for xylose. Up-regulation of purine biosynthesis genes was also detected in the transcriptome of *Escherichia coli* growing on glucose compared to xylose (Gonzalez *et al.*, 2002). On xylose, several gene clusters required for xylan or xylose conversion were up-regulated (Csac0692-0696; Csac0240-0242; Csac2416-2419). These clusters encode ABC transport systems, transcriptional regulators and endo-xylanases. In addition, genes, specifically required for growth on rhamnose, were highly up-regulated during growth on rhamnose, thus indicating the utilization pathway for this sugar.

Discussion

C. saccharolyticus has been shown to be an excellent candidate for biohydrogen production (de Vrije *et al.*, 2007; Kadar *et al.*, 2004; van Niel *et al.*, 2002). In contrast to mesophilic fermentative anaerobes, it produces almost no reduced end products, such as lactate or ethanol, and the amount of hydrogen approaches the “Thauer limit” of 4 H_2/glucose (Thauer, 1976). Moreover, *C. saccharolyticus* hydrolyzes various biomass-derived polymers, such as cellulose, hemicellulose, starch, and pectin, and ferments corresponding sugar monomers, including glucose and xylose. The complete genome sequence of *C. saccharolyticus* provides new insights into the exceptional capacity of the bacterium to degrade a variety of plant polysaccharides and further reveals its high plasticity with many transposases, CRISPRs and two uneven replication arms. A large number of sugar hydrolases and transferases could be identified, outnumbering those in the hyperthermophile *Thermotoga maritima* (Chhabra *et al.*, 2003). Metabolic pathways for the degradation of residual components of cellulose, hemicellulose, starch and pectin could be assigned. Reducing equivalents are produced as NADH or reduced ferredoxin, which are apparently used directly to produce hydrogen by a soluble NADH-dependent Fe-only hydrogenase and a membrane-bound ferredoxin-dependent Ni-Fe hydrogenase. The ability to produce hydrogen directly from NADH is not known from mesophilic anaerobes and may be responsible for the relatively high hydrogen production rates by *C. saccharolyticus*. In mesophiles, reducing equivalents from NADH first have to be transferred to ferredoxin, which requires input of energy, either by a sodium gradient (Boiangiu *et al.*, 2005) or by coupling to an exergonic reaction (Li *et al.*, 2007). In hyperthermophiles, such as *C. saccharolyticus*, this is apparently not necessary.

The absence of catabolite repression by glucose is an important characteristic for biohydrogen producers since it allows them to process an array of biomass-derived substrates simultaneously. Whereas glucose is generally known to repress the use of xylose by CCR (Hueck and Hillen, 1995), this was not observed in *C. saccharolyticus.* Moreover, the transcriptome showed that the various components of a CCR system, present in the genome (CcpA homologs, HPr, HPr kinase), were not differentially transcribed under the conditions examined, suggesting that this type of regulation is not triggered by glucose or xylose. No obvious differences were noted in the transcriptome for central metabolic pathways during growth on either glucose or xylose or a mixture of both. The EM pathway was not affected by the hexose or pentose substrate, which is in contrast to the transcriptome analysis of *E. coli* for growth on the same substrates (Gonzalez *et al.*, 2002). Remarkably, also the same specific ABC transporter is upregulated on both substrates, which is also in line with the non-preferential behavior of *C. saccharolyticus* towards these two monomeric sugars.

Detailed knowledge on the metabolic pathways leading to hydrogen production enables one to identify key enzymes that may be targets for improving the H_2 yield by metabolic engineering. Currently, a genetic system for *C. saccharolyticus* is under development which will initially target the dehydrogenases involved in lactate and ethanol formation. Alternatively, genes could be introduced to constitute an ox-PPP, to achieve higher H_2 yields greater than 4 per mole of glucose (Zhang *et al.*, 2007). In any case, the *C. saccharolyticus* genome provides new insights into the metabolic features of a versatile biohydrogen producer, which can inspire efforts to optimize microbial bioenergy systems.

Materials and methods

Cultivation and DNA isolation

C. saccharolyticus (DSM 8903/ATCC 43494) was cultured overnight on DSMZ 640 medium at 70 ºC with glucose (50 mM) as carbon and energy source. Cells were harvested and genomic DNA was isolated according to the method of (Pitcher *et al.*, 1989) using guanidinium thiocyanate. Residual protein was removed in an additional purification step with SDS and proteinase K, followed by chloroform/isoamylalcohol extraction and isopropanol precipitation.

Genome Sequencing and assembly

High molecular weight genomic DNA was provided to the US Department of Energy Joint Genome Institute (http://www.jgi.doe.gov/) for cloning and shotgun sequencing. A combination of small (average insert sizes: 3, 8 kb) and large (40 kb, fosmid) insert libraries were prepared and used for analysis as indicated at http://www.jgi.doe.gov/. The complete final assembly was released on 8-May-2007 and listed under GenBank accession #CP000679 (http://genome.jgi-psf.org/finished_microbes/calsa/calsa.home.html).

Genome annotation and comparative analysis

Critica (Badger and Olsen, 1999) and Glimmer (Delcher *et al.*, 1999) software programs were used for coding region detection and gene identification. TMMHMM 2.0 (Krogh *et al.*, 2001) was used to predict transmembrane helices in translated sequences. SignalP v2.0b2 (Nielsen *et al.*, 1997) was used to predict the presence and location of N-terminal signal peptides. All automatic gene and function predictions were manually checked using BLAST programs (Altschul *et al.*, 1990), InterPro (Mulder *et al.*, 2007) and The Integrated Microbial Genomes (IMG) system (Markowitz *et al.*, 2006) and corrected if necessary. Protein functions were checked with Carbohydrate-Active enzymes (CAZy; http://www.cazy.org (Coutinho and Henrissat, 1999)) classification. The comparative analysis was conducted based on the assignment and classification of Clusters of Orthologous Groups of proteins (COG) (Tatusov *et al.*, 2003) by the IMG system.

Growth experiments and RNA isolation

C. saccharolyticus was subcultured (overnight) 3 times on the substrate of interest in modified DSMZ 640 medium before inoculating a pH-controlled (pH = 7) 1-liter fermentor containing 4 gram substrate per liter. Cells were grown at 70 ºC until mid-logarithmic phase (~OD_{660} = 0.3-0.4) and harvested by centrifugation and rapid cooling to 4 ºC and stored at -80 ºC. Total RNA was isolated using a modified Trizol (Invitrogen) protocol in combination with an RNA easy kit (Qiagen). Quality was tested with the Experion Bioanalyzer (Biorad) and cDNA was constructed with Superscript III reverse transcriptase (Invitrogen).

Whole genome oligonucleotide DNA microarray design and construction

A DNA microarray was designed and constructed based on 2695 protein-coding sequences in the *C. saccharolyticus* genome obtained from the Department on Energy's Joint Genomes Institute (http://genome.ornl.gov/microbial/csac). OligoArray 2.0 (Rouillard *et al.*, 2003) was used to generate one 60-mer oligonucleotide probe sequence for each open reading frame. The probes were synthesized (Integrated DNA Technologies, IA), re-suspended in 50% DMSO, and printed onto Ultragap microarray slides (Corning, NY) using a QArrayMini arrayer (Genetix, UK). Each probe was spotted five times onto each array to fortify statistical analysis.

Microarray hybridization

The cDNA samples were processed using the Qiaquick purification kit (Qiagen, CA) with the cDNA samples eluted using phosphate buffer. The quantity and quality of the recovered cDNA samples were subsequently analyzed with absorbance at 260/280 nm. Cyanine-3 and Cyanine-5 dye (Amersham, UK) labeling and sample hybridizations were done following instructions from TIGR (http://www.tigr.org/tdb/microarray/protocolsTIGR.shtml), with minor adjustments to accommodate long-oligonucleotide platforms. Samples were hybridized in a 4-slide loop (Supplementary Fig. S5.1).

Data collection and analysis

After incubation, slides were washed to remove non-specifically bound material, and scanned with a ScanArray Lite microarray scanner (Perkin Elmer, MA). Data acquisition and spot quantitation were performed with the ScanArray Express software. Once all the slides were quantitated, data from the loop was analyzed with JMP Genomics 3.0 (SAS, NC), as described previously (Pysz *et al.*, 2004) using a mixed effects ANOVA model (Wolfinger *et al.*, 2001).

Acknowledgements

The authors gratefully acknowledge the US Department of Energy and the Joint Genome Institute for providing the genome sequence for Caldicellulosiruptor saccharolyticus. This research is supported by the Earth and Life Sciences Foundation (ALW) and Chemische Wetenschappen (CW) project 050.50.206 (subsidized by the Netherlands Organization for Scientific Research (NWO)) and by the European Commission (HYVOLUTION project; Reference nr. 19825). Microarray fabrication was financially supported in part by Danisco (Genencor International, Inc., Palo Alto, USA). RMK acknowledges support from the US NSF Biotechnology Program. ALV and DLL acknowledge support from a US Department of Education GAANN Fellowship.

Supplementary material

For supplementary tables and figure see Appendix I.

References

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

Badger, J. H., and Olsen, G. J. (1999). CRITICA: coding region identification tool invoking comparative analysis. Mol Biol Evol *16*, 512-524.

Bao, Q., Tian, Y., Li, W., Xu, Z., Xuan, Z., Hu, S., Dong, W., Yang, J., Chen, Y., Xue, Y*., et al.* (2002). A complete sequence of the *T. tengcongensis* genome. Genome Res *12*, 689-700.

Barrangou, R., Fremaux, C., Deveau, H., Richards, M., Boyaval, P., Moineau, S., Romero, D. A., and Horvath, P. (2007). CRISPR provides acquired resistance against viruses in prokaryotes. Science *315*, 1709-1712.

Barriere, C., Veiga-da-Cunha, M., Pons, N., Guedon, E., van Hijum, S. A., Kok, J., Kuipers, O. P., Ehrlich, D. S., and Renault, P. (2005). Fructose utilization in *Lactococcus lactis* as a model for low-GC gram-positive bacteria: its regulator, signal, and DNA-binding site. J Bacteriol *187*, 3752-3761.

Bergquist, P. L., Gibbs, M. D., Morris, D. D., Te'o, V. S. J., Saul, D. J., and Morgan, H. W. (1999). Molecular diversity of thermophilic cellulolytic and hemicellulolytic bacteria. FEMS Microbiol Ecol *28*, 99-110.

Boiangiu, C. D., Jayamani, E., Brugel, D., Herrmann, G., Kim, J., Forzi, L., Hedderich, R., Vgenopoulou, I., Pierik, A. J., Steuber, J., and Buckel, W. (2005). Sodium ion pumps and hydrogen production in glutamate fermenting anaerobic bacteria. J Mol Microbiol Biotechnol *10*, 105-119.

Boyd, D. A., Cvitkovitch, D. G., and Hamilton, I. R. (1995). Sequence, expression, and function of the gene for the nonphosphorylating, NADP-dependent glyceraldehyde-3-phosphate dehydrogenase of *Streptococcus* mutans. J Bacteriol *177*, 2622-2627.

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

Claassen, P. A. M., and de Vrije, T. (2006). Non-thermal production of pure hydrogen from biomass: HYVOLUTION. International Journal of Hydrogen Energy *31*, 1416-1423.

Conners, S. B., Montero, C. I., Comfort, D. A., Shockley, K. R., Johnson, M. R., Chhabra, S. R., and Kelly, R. M. (2005). An expression-driven approach to the prediction of carbohydrate transport and utilization regulons in the hyperthermophilic bacterium *Thermotoga maritima*. J Bacteriol *187*, 7267-7282.

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

de Vrije, T., Mars, A. E., Budde, M. A., Lai, M. H., Dijkema, C., de Waard, P., and Claassen, P. A. (2007). Glycolytic pathway and hydrogen yield studies of the extreme thermophile *Caldicellulosiruptor saccharolyticus*. Appl Microbiol Biotechnol *74*, 1358-1367.

Delcher, A. L., Harmon, D., Kasif, S., White, O., and Salzberg, S. L. (1999). Improved microbial gene identification with GLIMMER. Nucleic Acids Res *27*, 4636-4641.

EC2002 (2002). Future needs and challenges for non-nuclear energy research in the European Union. Discussion Paper.

Gonzalez, R., Tao, H., Shanmugam, K. T., York, S. W., and Ingram, L. O. (2002). Global gene expression differences associated with changes in glycolytic flux and growth rate in *Escherichia coli* during the fermentation of glucose and xylose. Biotechnol Prog *18*, 6-20.

Hueck, C. J., and Hillen, W. (1995). Catabolite repression in *Bacillus subtilis*: a global regulatory mechanism for the gram-positive bacteria? Mol Microbiol *15*, 395-401.

Hyun, H. H., Shen, G. J., and Zeikus, J. G. (1985). Differential amylosaccharide metabolism of *Clostridium thermosulfurogenes* and *Clostridium thermohydrosulfuricum*. J Bacteriol *164*, 1153-1161.

Kadar, Z., de Vrije, T., van Noorden, G. E., Budde, M. A., Szengyel, Z., Reczey, K., and Claassen, P. A. (2004). Yields from glucose, xylose, and paper sludge hydrolysate during hydrogen production by the extreme thermophile *Caldicellulosiruptor saccharolyticus*. Appl Biochem Biotechnol *113-116*, 497-508.

Kerkhoven, R., van Enckevort, F. H., Boekhorst, J., Molenaar, D., and Siezen, R. J. (2004). Visualization for genomics: the Microbial Genome Viewer. Bioinformatics *20*, 1812-1814.

Kikuchi, A., and Asai, K. (1984). Reverse gyrase--a topoisomerase which introduces positive superhelical turns into DNA. Nature *309*, 677-681.

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

Li, F., Hinderberger, J., Seedorf, H., Zhang, J., Buckel, W., and Thauer, R. K. (2007). Coupled ferredoxin- and crotonyl-CoA reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from *Clostridium kluyveri*. J Bacteriol.

Ludwig, H., Homuth, G., Schmalisch, M., Dyka, F. M., Hecker, M., and Stulke, J. (2001). Transcription of glycolytic genes and operons in *Bacillus subtilis*: evidence for the presence of multiple levels of control of the *gapA* operon. Mol Microbiol *41*, 409-422.

Markowitz, V. M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I*., et al.* (2006). The integrated microbial genomes (IMG) system. Nucleic Acids Res *34*, D344-348.

Miwa, Y., Nakata, A., Ogiwara, A., Yamamoto, M., and Fujita, Y. (2000). Evaluation and characterization of catabolite-responsive elements (*cre*) of *Bacillus subtilis*. Nucleic Acids Res *28*, 1206-1210.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R*., et al.* (2007). New developments in the InterPro database. Nucleic Acids Res *35*, D224-228.

Nelson, K. E., Clayton, R. A., Gill, S. R., Gwinn, M. L., Dodson, R. J., Haft, D. H., Hickey, E. K., Peterson, J. D., Nelson, W. C., Ketchum, K. A*., et al.* (1999). Evidence for lateral gene transfer between Archaea and bacteria from genome sequence of *Thermotoga maritima*. Nature *399*, 323-329.

Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997). Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng *10*, 1-6.

Nieto, C., Espinosa, M., and Puyet, A. (1997). The maltose/maltodextrin regulon of *Streptococcus pneumoniae*. Differential promoter regulation by the transcriptional repressor MalR. J Biol Chem *272*, 30860-30865.

Pitcher, D. G., Saunders, N. A., and Owen, R. J. (1989). Rapid extraction of bacterial genomic DNA with guanidium thiocyanate. Lett Appl Microbiol *8*, 151-156.

Pysz, M. A., Ward, D. E., Shockley, K. R., Montero, C. I., Conners, S. B., Johnson, M. R., and Kelly, R. M. (2004). Transcriptional analysis of dynamic heat-shock response by the hyperthermophilic bacterium *Thermotoga maritima*. Extremophiles *8*, 209-217.

Rainey, F. A., Donnison, A. M., Janssen, P. H., Saul, D., Rodrigo, A., Bergquist, P. L., Daniel, R. M., Stackebrandt, E., and Morgan, H. W. (1994). Description of *Caldicellulosiruptor saccharolyticus* gen. nov., sp. nov: an obligately anaerobic, extremely thermophilic, cellulolytic bacterium. FEMS Microbiol Lett *120*, 263-266.

Ridley, B. L., O'Neill, M. A., and Mohnen, D. (2001). Pectins: structure, biosynthesis, and oligogalacturonide-related signaling. Phytochemistry *57*, 929-967.

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

Rouillard, J. M., Zuker, M., and Gulari, E. (2003). OligoArray 2.0: design of oligonucleotide probes for DNA microarrays using a thermodynamic approach. Nucleic Acids Res *31*, 3057-3062.

Sato, T., Imanaka, H., Rashid, N., Fukui, T., Atomi, H., and Imanaka, T. (2004). Genetic evidence identifying the true gluconeogenic fructose-1,6-bisphosphatase in *Thermococcus kodakaraensis* and other hyperthermophiles. J Bacteriol *186*, 5799-5807.

Soboh, B., Linder, D., and Hedderich, R. (2004). A multisubunit membrane-bound [NiFe] hydrogenase and an NADH-dependent Fe-only hydrogenase in the fermenting bacterium *Thermoanaerobacter tengcongensis*. Microbiology *150*, 2451-2463.

Stams, A. J. (1994). Metabolic interactions between anaerobic bacteria in methanogenic environments. Antonie Van Leeuwenhoek *66*, 271-294.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N*., et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

Te'o, V. S., Saul, D. J., and Bergquist, P. L. (1995). *celA*, another gene coding for a multidomain cellulase from the extreme thermophile *Caldocellum saccharolyticum*. Appl Microbiol Biotechnol *43*, 291-296.

Thauer, R. (1976). Limitation of microbial hydrogen formation via fermentation. In Microbial Energy Conversion, H. G. Schlegel, and J. Barnea, eds. (Göttingen, Erich Goltze KG), pp. 201-294.

USDOE (2002). A National Vision of America's Transition to a Hydrogen Economy—to 2030 and Beyond

van Niel, E. W. J., Budde, M. A. W., de Haas, G. G., van der Wal, F. J., Claasen, P. A. M., and Stams, A. J. M. (2002). Distinctive properties of high hydrogen producing extreme thermophiles, *Caldicellulosiruptor saccharolyticus* and *Thermotoga elfii*. International Journal of Hydrogen Energy *27*, 1391-1398.

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

Zhang, Y. H., Evans, B. R., Mielenz, J. R., Hopkins, R. C., and Adams, M. W. (2007). High-yield hydrogen production from starch and water by a synthetic enzymatic pathway. PLoS ONE *2*, e456.

Chapter 6

General discussion and summary

This thesis describes the research that was conducted on genomic data of (hyper)thermophilic microorganisms with an emphasis on regulatory pathways and carbohydrate metabolism using a bioinformatics approach. Hyperthermophiles and thermophiles are able to grow optimally at temperatures above 80 ° and 60 °C, respectively. Among these (hyper)thermophiles both aerobic (*Sulfolobus* spp.) and anaerobic (*Pyrococcus furiosus*, *Thermococcus kodakaraensis*, *Caldicellulosiruptor saccharolyticus)* representatives can be found.

Since the first completely sequenced genome of a free living organism *Haemophilus influenza* in 1995 (Fleischmann *et al.*, 1995), the number of complete genomes available has increased tremendously. At the time of writing, the complete genome sequence of 713 unicellular and multicellular organisms is known (Liolios *et al.*, 2007) and over 2500 sequencing projects are ongoing. Computational genomics is a key component in assembling and analyzing the genomes generated by genome sequencing projects. In this thesis computational analyses were carried out on the completely sequenced genomes of the hyperthermophilic archaea *P. furiosus*, *T. kodakaraensis* and *S. solfataricus* and the extremely thermophilic bacterium *C. saccharolyticus* (Table 6.1); in addition, annotation was conducted on two plasmids of '*Sulfolobus islandicus*' SOG2/4.

Table 6.1 Completely sequenced genomes of hyperthermophiles and thermophiles in the Genomes Online Database (GOLD) (Liolios *et al.*, 2007) and are profoundly analyzed in this thesis.

Species	**Strain**	**Lifestyle**		**T_{opt} (°C)[a]**	**Genome size (kbp)**	**Proteins**	**GC-content (%)**	**Chromosomes**	**Plasmids**	**Reference**
Archaea										
Pyrococcus furiosus	JCM 8422	AN	H	100	1908	2125	40.8	1	0	(Robb *et al.*, 2001)
Sulfolobus solfataricus	P2	AE	H	80	2992	2977	35.8	1	0	(She *et al.*, 2001)
Thermococcus kodakaraensis	KOD1	AN	H	85	2088	2306	52	1	0	(Fukui *et al.*, 2005)
Bacteria										
Caldicellulosiruptor saccharolyticus	DSM 8903	AN	H	70	2970	2679	35	1	0	**Chapter 5**

AE, aerobe; AN, anaerobe; H, heterotroph.

[a] $T_{opt:}$ optimal growth temperature or temperature growth range of microbes according to GOLD database, Prokaryotic Growth Temperature database (PGTdb) (Huang *et al.*, 2004) or species description.

This thesis employs a three-pillar approach: Computational genomics, functional genomics and classical molecular biology (Fig. 6.1). The computational pillar generates hypotheses based on high-throughput genome analyses, which should be tested by either high-throughput post-genome analysis, or by classical experimental analysis. The high-throughput functional genomics tools include transcriptomics (DNA microarrays), proteomics (*e.g.* by 2D electrophoresis in combination with LC-MS) and metabolomics (*e.g.* by NMR, FT-MS). These analyses are generally used to verify *in silico* predictions produced by comparing different species, different

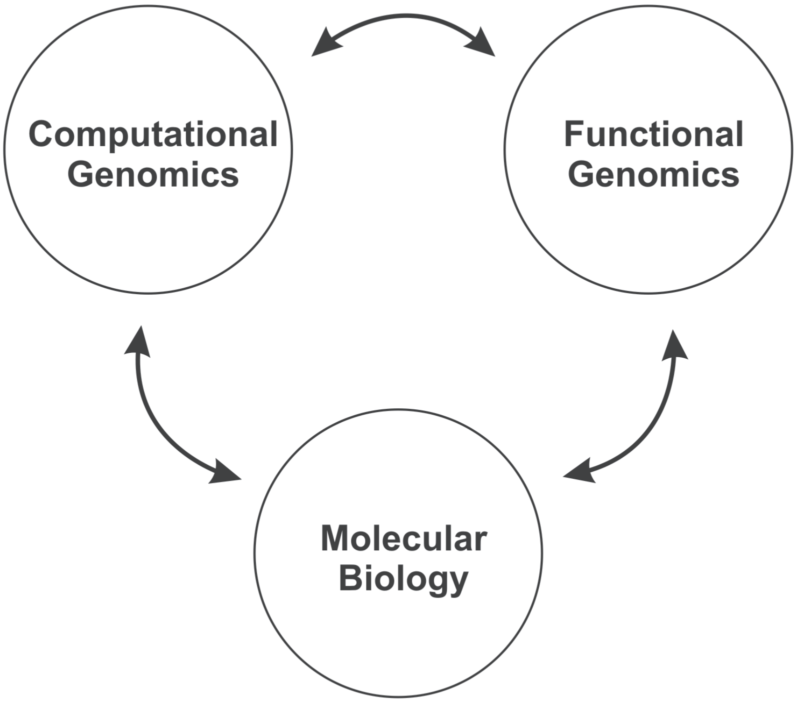


Figure 6.1 Inter-disciplinary approach as described in this thesis through integration of the three pillars: the two high-throughput approaches computational genomics and functional genomics, as well as the classical molecular biology approach that includes genetics, biochemistry and physiology.

cultivation conditions of a single species, or by comparing different genotypes (wild type vs. mutant). The hypotheses generated are to be confirmed by classical molecular biology experiments (genetics, biochemistry and physiology) (Fig. 6.1). The three pillars should not operate independently, but rather form the basis for building a biological model, that by continuous integration in an iterative way should lead to new biological insights. Such integrated genomics approaches have been used in this thesis to study several hyperthermophilic organisms, with the main focus on unraveling central carbohydrate metabolism, enzymes, pathways and their control. Table 6.2 gives an overview of the chapters, tools and species described in this thesis, as well as reported elsewhere.

Table 6.2 Microbial genomes analyzed and explored in thesis and in additional publications. The research is divided into three distinct categories: (1) computational genomics, (2) functional genomics and (3) molecular biology including classical biochemistry, genetics and physiology.

Species	Computational genomics	Functional genomics	Molecular biology
Sulfolobus spp.	**Chapter 2** **Chapter 3** (Brouns *et al.*, 2006) (Erauso *et al.*, 2006)	**Chapter 2** (Brouns *et al.*, 2006)	(Brouns *et al.*, 2006)
Pyrococcus furiosus/ Thermococcus kodakaraensis	**Chapter4**	(Kanai *et al.*, 2007)	**Chapter 4** (Kanai *et al.*, 2007)
Caldicellulosiruptor saccharolyticus	**Chapter 5**	**Chapter 5**	**Chapter 5**

Metabolic reconstruction of carbohydrate metabolism in *Sulfolobus* spp.

A computational and functional genomics approach was applied on the model crenarchaeon *S. solfataricus.* The response of *S. solfataricus* to different carbon sources, glucose versus tryptone and yeast extract, was studied. The complete transcriptome was analyzed and a 2D-electrophoresis map was reconstructed. In addition, ^{15}N-labeling technique was used to detect, at the protein level, the differentially expressed genes of the reconstructed central carbon metabolism. Remarkably, only three genes (14%) showed a clear regulatory profile (**Chapter 2**). This situation differs significantly from a comparable study in *Pyrococcus furiosus* (Schut *et al.*, 2003) and *Thermococcus kodakaraensis* (Kanai *et al.*, 2007) where extensive transcriptional regulation of the glycolytic genes was observed.

A similar approach as in **Chapter 2** (*S. solfataricus grown on* D-arabinose vs. D-glucose) has recently revealed a novel D-arabinose degradation pathway (Brouns *et al.*, 2006). After a comprehensive comparative, transcriptome and proteome analysis was carried out, a general prokaryotic pentose, hexaric acid and hydroxyl-l-proline catabolic pathway was proposed that ends in α-keto-glutarate, a component of the citric acid cycle. A putative *cis*-regulatory element (ARA-box) was predicted and is most likely involved in transcriptional regulation of arabinose in *S. solfataricus*. Moreover, putative functions of the enzymes were confirmed biochemically and can be copied to enzymes in a wide-range of the aforementioned prokaryotic pathways. Therefore, the literature on the diversity of the anabolic, catabolic pentose metabolism and regulatory mechanisms was reviewed (**Chapter 3**). This review comprised the metabolic reconstruction of the pentose utilizing pathways in Archaea and these reconstructions were compared to Bacteria and Eukarya. The enzymes involved in the pathways did probably evolve by recruitment events and exemplify the existence of a 'variable metabolic shell' in addition to a 'conserved housekeeping core' in prokaryotes, which allows adjusting the metabolic infrastructure to available substrates. However, the regulatory mechanisms are still unknown and therefore an interesting area for future research, in particular with the new transcriptomics and proteomics tools.

Transcriptional regulation of metabolic pathways

A promoter analysis tool was developed to analyze small regulatory elements that are vital in the regulation of the transcription of genes. The analysis of glycolytic genes in the anaerobic species of the order Thermococcales and the primer extension analysis in *P. furiosus* pointed out a clear palindromic candidate that controls all the genes of the glucose and starch-degrading pathway. This Thermococcales-Glycolytic-Motif (TGM) seemed to be involved in controlling the pathways of glucose construction and degradation. Furthermore, a comparative

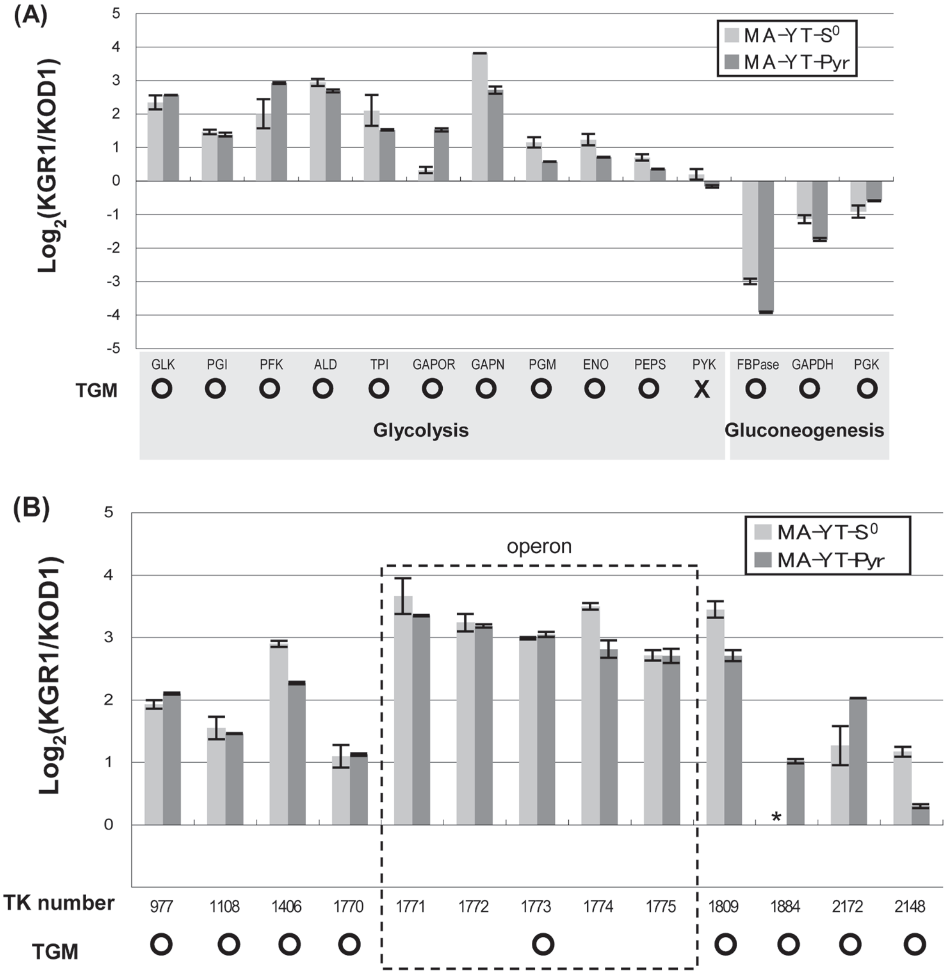


Figure 6.2 **(A)** Relative transcript levels of selected genes related to glycolysis and gluconeogenesis in KOD1 and KGR1 cells grown under gluconeogenic (MA-YT-S0 or MA-YT-Pyr) conditions. **(B)** Relative transcript levels of selected genes related to maltodextrin metabolism in KOD1 and KGR1 cells grown under gluconeogenic conditions (MA-YT-S0 or MA-YT-Pyr). The presence (O) or absence (X) of a TGM in each promoter is indicated and a *asterisk* indicates that the datum was not reliable because of low signal intensity. *Error bars* were calculated from two sets of microarrays. *GLK*, ADP-dependent glucokinase (TK1110); *PGI*, glucose-6-phosphate isomerase (TK1111); *PFK*, ADP-dependent PFK (TK0376); *ALD*, fructose-1,6-bisphosphate aldolase (TK0989); *TPI*, triose-phosphate isomerase (TK2129); *GAPOR*, GAP:ferredoxin oxidoreductase (TK2163);*GAPN*,GAPdehydrogenase (non-phosphorylating) (TK0705); *PGM*, phosphoglycerate mutase (TK0866); *ENO*, enolase (TK2106); *PEPS*, phosphoenolpyruvate synthase (TK1292); *PYK*, pyruvate kinase (TK0511); *FBPase*, TK2164; *GAPDH*, GAP dehydrogenase (phosphorylating) (TK0765); *PGK*, 3-phosphoglycerate kinase (TK1146). *B*, relative transcript levels of selected genes related to maltodextrin metabolism under glycolytic (MA-YT-Mdx) and gluconeogenic conditions (MA-YT-S0 or MA-YTPyr) in the wild-type strain. Genes constituting a putative operon are *boxed* with a *dotted line*. *TK0977*, pullulanase type II, GH13 family; *TK1108*, phosphohexomutase; *TK1406*, maltodextrin phosphorylase; *TK1770*, cyclomaltodextrinase; *TK1771*, maltodextrin-binding protein precursor; *TK1772*, maltodextrin transport system, permease component; *TK1773*, maltodextrin transport system, permease component; *TK1774*, amylopullulanase; *TK1775*, maltodextrin transport system, ATPase component; *TK1809*, 4-α-glucanotransferase; *TK1884*, α-amylase; *TK2148*, α-glucosidase; *TK2172*, cyclomaltodextrin glucanotransferase. This figure was previously published in (Kanai *et al.*, 2007).

genomic analysis of the hyperthermophilic Thermococcales species, some of which have the ability to degrade starch, revealed a transcriptional regulator, a homolog of the characterized transcriptional regulator of *mal* operon -TrmB - (Lee *et al.*, 2003). On the basis of this analysis, it has been hypothesized that this "glycolytic" regulator might be the transcriptional regulator involved in binding the TGM promoter motif in the absence of starch (**Chapter 4**). The putative regulator of the largest predicted archaeal regulon to date has been tested by a novel knock-out system in *Thermococcus kodakaraensis* (Kanai *et al.*, 2007). The Thermococcales glycolytic regulator Tgr (TK1769) was disrupted and the strain (Δ*tgr*) indeed showed the predicted phenotype: an impaired growth rate under gluconeogenic conditions *vs*. the wild type. A whole genome transcriptome analysis showed relatively high levels of transcripts of almost all genes

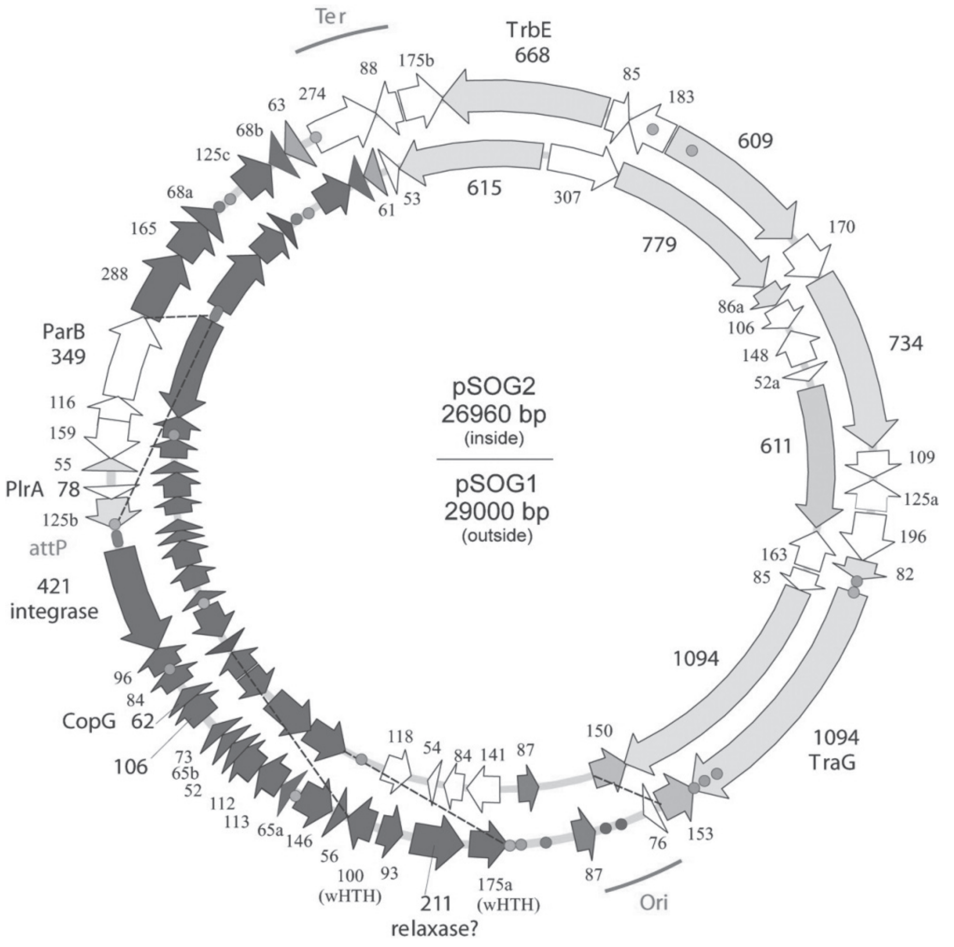


Figure 6.3 (for color figure see Appendix II) Comparison of the pSOG1 and pSOG2 sequences. This diagram shows the circular genomes of pSOG1 on the outside and pSOG2 on the inside. ORFs are shown as arrows. Similar ORFs in the two plasmids are filled in gray; identical ORFs are filled in black; ORFs not conserved between the two plasmids are not filled. ORFs with predicted functions are labelled and ORFs discussed in the text are in bold. Insertions and gene replacements are indicated by dashed lines between the two genomes. ORF names are shown next to the corresponding arrows. The recombination motif TAAACTGGGGAGTTTA is represented by a small disk, colored green when present on the direct DNA strand and light blue when located on the complementary strand. Blue disks indicate the two larger tandem repeats, and a red disk indicates larger inverted repeats. The violet oval represents the putative site of integration attP. The approximate location of the origin (Ori) and terminus (Ter) of replication as predicted by cumulative GC skew and Z-curve analyses are also indicated. This figure was previously published in (Erauso *et al.*, 2006).

related to glycolysis and maltodextrin metabolism (Fig. 6.2A/B). In addition, the strain (Δ*tgr*) displayed a transcriptional activation defect of gluconeogenic genes (Fig. 6.2A). Besides the *in vitro* confirmation of the interaction of the Tgr and the TGM by electrophoretic mobility shift assay and the subsequent release of the regulator due to the ligand maltotriose, the data clearly provide *in vivo* confirmation of the computational analysis.

Genome sequence analysis

Two conjugative plasmids (CPs) were isolated and characterized from '*Sulfolobus islandicus*' strain SOG2/4, and these plasmids were sequenced and a comparative analysis was performed with other *Sulfolobus* plasmids (Erauso *et al.*, 2006). The comparative analysis showed a well-conserved core and revealed that 70% of the plasmids pSOG1 and pSOG2 is different (Fig. 6.3). The differences consist of a mixture of genes that often resemble counterparts in previously described *Sulfolobus* CPs.

Finally, the genome of the extremely thermophilic bacterium Caldicellulosiruptor saccharolyticus was completely sequenced and annotated. Moreover, a complete transcriptome analysis of C. saccharolyticus, grown on a variety of monosaccharides, was performed. This extraordinary organism is able to degrade cellulose, hemicellulose and several other poly- and monosaccharides and generates substantial amounts of hydrogen as fermentation end product. Biological H2 production is seen as one of the options for renewable H2 production on the longer term. The genome sequence shows similarities with the phylogenetically close clostridia and distantly related hyperthermophiles. Along with the expression data, the genomic sequence reveals the capability of C. saccharolyticus to grow on many different sugars as well as the transcriptional regulation of genes involved in the breakdown of these carbohydrates (Chapter 5). The extraordinary property of C. saccharolyticus to simultaneously metabolize pentoses and hexoses was displayed and the hydrogen yield of almost 4 H2 per mol of glucose was confirmed.

Discussion and future perspectives

Hyperthermophilic organisms have been discovered since the early 1970s. The discovery of these extremophilic microbes has opened a new unexplored field in microbiology that contributes to fundamental insights in physiology, molecular biology and biochemistry. Additionally, proteins from hyperthermophilic hosts have been proven to be applicable in the industry and molecular biology (Vieille and Zeikus, 2001), and employing the whole cell as cell factory is certainly feasible. However, isolating and obtaining a pure culture of microorganism, and in particular hyperthermophiles, is still not straight-forward. Although, new tools have been developed including: the use of phylogenetic staining and 'optical tweezers' (Huber *et al.*, 1995) or the use of microbial chips (Ingham *et al.*, 2007).

The cultivation and the purification of a microorganism provide the possibility to sequence the complete genome of one single organism. Since the development of the nucleotide sequencing technique by Frederick Sanger (Sanger and Coulson, 1975) much progress has been made at the level of optimizing this technology. The technological progress has resulted in a dramatic increase of the throughput of DNA-sequencing, and a substantial cost reduction. As a consequence, the number of genome sequencing projects is still growing exponentially to a number of 3557 at the time of writing (Liolios *et al.*, 2007). At present the new sequencing technique (pyrosequencing) (Ronaghi *et al.*, 1998) and the Solexa/Illumina technique will not only lead to cheaper genome sequencing projects, but will also give the opportunity to quantify the transcriptome by sequencing cDNA. In addition, metagenomics, *i.e.* the study of genetic material in an environment, will be economically feasible for hyperthermophilic environments. However, these new techniques require new algorithms and much more computational power to analyze the high-throughput data.

Although sequencing complete genome is getting easier and more cost-effective, and in spite of the fact that the prediction of protein-coding genes in prokaryotes is currently very accurate, the actual function prediction is still a major problem. Improved algorithms to predict homologous and orthologous relationships among proteins; improved data repositories for physiological data, metabolic pathways, expression data and protein-protein interactions and; enhanced accessibility of these database systems; would improve the quality and will accelerate gene and genome annotations. In addition, innovations in genetics and biochemistry, which could clarify protein functions, can have a significant contribution to genome annotation. The accurate promoter prediction in *Pyrococcus* and *Thermococcus* is a very good example of illustrating the power of *in silico* analysis. In genetics and computational genomics new developments are emerging on hyperthermophiles, particularly in Archaea (Allers and Mevarech, 2005; Makarova *et al.*, 2007). These new developments can be used to engineer, for instance, the metabolism of *S. solfataricus* for the production of biofuels (Blanch *et al.*, 2008). Additionally, a genetic system for the H_2 producing bacterium *C. saccharolyticus* is currently being developed **(Chapter 5)**. Moreover, at the level of functional genomics, as described above, the ongoing exponential growth of new genomic sequences releases do require even more efficient bioinformatics analysis tools.

Finally, it is important to emphasize that the reductive approach, which has been quite successful in understanding certain details of the biology of the hyperthermophilic cell, has been criticized. The analysis of isolated parts of a microbial cell will not give a complete picture of the dynamics and structure of the system. Systems biology is a holistic approach, aiming at a complete interpretation of what is happening in the microbial cell, by trying to fully uncover the interactions and emerging properties of the cell (Kitano, 2002). Systems biology is not only integrating data from different kind of experiments, such as proteomic, transcriptomic and metabolomic data, but also quantifying biomolecules and their interactions by computational modeling. Computational biology is therefore an important factor in studying biology, and will be even more important in the future.

References

Allers, T., and Mevarech, M. (2005). Archaeal genetics - the third way. Nat Rev Genet *6*, 58-73.

Blanch, H. W., Adams, P. D., Andrews-Cramer, K. M., Frommer, W. B., Simmons, B. A., and Keasling, J. D. (2008). Addressing the need for alternative transportation fuels: the Joint BioEnergy Institute. ACS Chem Biol *3*, 17-20.

Brouns, S. J., Walther, J., Snijders, A. P., van de Werken, H. J., Willemen, H. L., Worm, P., de Vos, M. G., Andersson, A., Lundgren, M., Mazon, H. F.*, et al.* (2006). Identification of the missing links in prokaryotic pentose oxidation pathways: evidence for enzyme recruitment. J Biol Chem *281*, 27378-27388.

Erauso, G., Stedman, K. M., van de Werken, H. J., Zillig, W., and van der Oost, J. (2006). Two novel conjugative plasmids from a single strain of Sulfolobus. Microbiology *152*, 1951-1968.

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

Fukui, T., Atomi, H., Kanai, T., Matsumi, R., Fujiwara, S., and Imanaka, T. (2005). Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with *Pyrococcus* genomes. Genome Res *15*, 352-363.

Huang, S. L., Wu, L. C., Liang, H. K., Pan, K. T., Horng, J. T., and Ko, M. T. (2004). PGTdb: a database providing growth temperatures of prokaryotes. Bioinformatics *20*, 276-278.

Huber, R., Burggraf, S., Mayer, T., Barns, S. M., Rossnagel, P., and Stetter, K. O. (1995). Isolation of a hyperthermophilic archaeum predicted by in situ RNA analysis. Nature *376*, 57-58.

Ingham, C. J., Sprenkels, A., Bomer, J., Molenaar, D., van den Berg, A., van Hylckama Vlieg, J. E., and de Vos, W. M. (2007). The micro-Petri dish, a million-well growth chip for the culture and high-throughput screening of microorganisms. Proc Natl Acad Sci U S A *104*, 18217-18222.

Kanai, T., Akerboom, J., Takedomi, S., van de Werken, H. J., Blombach, F., van der Oost, J., Murakami, T., Atomi, H., and Imanaka, T. (2007). A global transcriptional regulator in Thermococcus kodakaraensis controls the expression levels of both glycolytic and gluconeogenic enzyme-encoding genes. J Biol Chem *282*, 33659-33670.

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

Lee, S. J., Engelmann, A., Horlacher, R., Qu, Q., Vierke, G., Hebbeln, C., Thomm, M., and Boos, W. (2003). TrmB, a sugar-specific transcriptional regulator of the trehalose/maltose ABC transporter from the hyperthermophilic archaeon *Thermococcus litoralis*. J Biol Chem *278*, 983-990.

Liolios, K., Mavromatis, K., Tavernarakis, N., and Kyrpides, N. C. (2007). The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. Nucleic Acids Res.

Makarova, K. S., Sorokin, A. V., Novichkov, P. S., Wolf, Y. I., and Koonin, E. V. (2007). Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol Direct *2*, 33.

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

Ronaghi, M., Uhlen, M., and Nyren, P. (1998). A sequencing method based on real-time pyrophosphate. Science *281*, 363, 365.

Sanger, F., and Coulson, A. R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol *94*, 441-448.

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

She, Q., Singh, R. K., Confalonieri, F., Zivanovic, Y., Allard, G., Awayez, M. J., Chan-Weiher, C. C., Clausen, I. G., Curtis, B. A., De Moors, A., *et al.* (2001). The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc Natl Acad Sci U S A *98*, 7835-7840.

Vieille, C., and Zeikus, G. J. (2001). Hyperthermophilic enzymes: sources, uses, and molecular mechanisms for thermostability. Microbiol Mol Biol Rev *65*, 1-43.

Nederlandse Samenvatting

Nederlandse samenvatting

Dit proefschrift beschrijft de rekenkundige analyse van het gehele erfelijk materiaal, ook wel genoom genoemd, van organismen die optimaal groeien boven temperaturen van 80 ºC (hyperthermofielen) of 60 ºC (thermofielen). Deze warmteminnende organismen zijn allen eencellig en hebben geen celkern; de zogenaamde prokaryoten. Deze prokaryoten zelf zijn weer in twee fundamenteel verschillende domeinen in te delen: het domein van de Bacteriën en van de Archaea. Het doel van het onderzoek beschreven in dit proefschrift is het begrijpen hoe het genoom in deze organismen functioneert. De rekenkundige analyses hebben zich toegespitst op het in kaart brengen van genen, het voorspellen van de functies van eiwitten, het ontrafelen van metabole routes, het verkrijgen van inzicht in het proces van genexpressie en het beschrijven van evolutie van prokaryoten. Tevens zijn met behulp van bioinformatica grote datasets met genexpressie data van eiwit en boodschapper-RNA geanalyseerd en zijn de analyses getoetst met klassieke microbiologische en moleculair biologische gegevens en experimenten. Het onderzoek heeft zich toegespitst op de koolhydraatafbraak routes en hun regulatie bij vier verschillende (hyper)thermofielen. (1) *Sulfolobus solfataricus*, een archaeon dat gevonden kan worden op het land in zure heetwaterbronnen (solfataren). (2) *Pyrococcus furiosus*; en zijn mede-orde genoot (3) *Thermococcus kodakaraensis*. Beide organismen leven zuurstofloos. *P. furiosus* is gevonden in de zee bij het Italiaanse Vulcano eiland, terwijl *T. kodakaraensis* uit een solfatare op het Japanse Kodakara eiland is verkregen. Als laatste is uitvoerig de waterstof producerende (4) *Caldicellulosiruptor saccharolyticus* bestudeerd. Dit organisme, dat van een houtstronk uit een heetwaterbron is geïsoleerd, bevat eigenschappen die zeer nuttig kunnen zijn om waterstof, de energiedrager van de toekomst, te produceren uit plantenmaterialen.

In **hoofdstuk 1** wordt een overzicht gegeven van de karakteristieken van de hyperthermofielen waarvan de DNA-volgorde volledig bekend is. Daarnaast wordt er dieper in gegaan op het koolhydraatmetabolisme, de transcriptie regulatie en de fylogenie van deze fascinerende levensvormen. Buiten een overzicht van eigenschappen van deze eencelligen worden er rekenkundige en functionele technieken beschreven die toepast zijn om meer inzicht te krijgen in de eigenschappen van hyperthermofielen. Het accent ligt hierbij op het achterhalen van eiwitfunctie, het reconstrueren van metabole routes en de regulatie van genexpressie op het niveau van RNA.

Hoofdstuk 2 beschrijft de reconstructie van het centrale koolhydraatmetabolisme in *S. solfataricus*, gebaseerd op het volledig in kaart gebrachte genoom. Tevens zijn zowel de hoeveelheden RNA als de hoeveelheden eiwit van *S. solfataricus,* groeiend op peptiden of glucose, genoom-breed gemeten en vergeleken. Tot onze grote verassing waren de verschillen van genexpressie op beide niveau’s zeer klein en dus zeer verschillend vergeleken met andere hyperthermofielen zoals *P. furiosus*. Slechts drie enzymen hadden een verschillend expressie profiel. Een mogelijke verklaring is dat *S. solfataricus* de metabole routes reguleert op allosteer niveau dan wel met behulp van post-translationele modificaties.

Een zelfde aanpak als in **hoofdstuk 2** waarbij *S. solfataricus* op twee aparte koolstofbronnen (D-arabinose t.o.v. D-glucose) werd gegroeid, toonde een nieuwe metabole route in dit archaeon aan. Met genoom vergelijkingen en een volledig transcriptoom onderzoek zijn er vele routes opgehelderd die betrokken zijn bij suiker afbraak in zowel in *S. solfataricus* als in andere prokaryoten. Deze resultaten gaven hernieuwde inzichten in pentose metabolisme van Archaea. Het metabolisme van pentoses is opnieuw geanalyseerd en een uitgebreid literatuur onderzoek hiervan is gepresenteerd in **hoofdstuk 3**.

In **hoofdstuk 4** is de genexpressie in *P. furiosus* and *T. kodakaraensis* bestudeerd, waarvan voornamelijk de expressie van de genen van de afbraakroute van glucose (glycolyse). Met behulp van een zelfontwikkeld promoter-analyse-programma en de identificatie van de transcriptie start sites in de glycolytische genen is, naast de identificatie van een TATA-box en een TFB-responsive element (BRE), een nieuw *cis*-regulatory element ontdekt. Dit palindromisch motief is Thermococcales-Glycolytic-Motif (TGM) genoemd. TGM ligt voor alle genen van de glycolyse, maar ook voor genen die betrokken zijn bij afbraak en transport van grote ketens van glucose moleculen (maltodextrines en zetmeel) en bij de opbouw van glucose. Door de genomen van de twee bovengenoemde organismen te vergelijken met twee soortgenoten die geen zetmeel afbreken (*Pyrococcus abyssi* en *Pyrococcus horikoshii*) is er een kandidaat regulator gevonden (PF0124, TK1769) die het TGM kan binden. Deze regulator is verantwoordelijk voor het uit- of aanzetten van meer dan 20 genen.

In **hoofdstuk 5** wordt het gehele genoom van de extreem thermofiel *Caldicellulosiruptor saccharolyticus* beschreven. Het genoom bevat een circulair chromosoom bestaande uit bijna 3 miljoen basenparen. Het gehele genoom codeert voor 2.679 voorspelde eiwitten. De eiwitten worden door *C. saccharolyticus* o.a. gebruikt om op een veelheid van koolhydraten te groeien en hieruit waterstof te produceren. Naast het genoom is er ook een compleet transcriptoom analyse uitgevoerd, waarbij *C. saccharolyticus* gegroeid werd op verschillende monosachariden (D-rhamnose, D-glucose, D-xylose en een mengsel van D-glucose en D-xylose). Deze analyse toonde aan dat *C. saccharolyticus* de belangrijkste plantensuikers D-glucose en D-xylose gelijktijdig verbrand. Door deze eigenschap kan *C. saccharolyticus* vele soorten biomassa tegelijkertijd omzetten wat belangrijk is voor de productie van waterstof uit plantenmaterialen.

Hoofdstuk 6 bediscussieert de bevindingen uit dit proefschrift, vergelijkt het met recente inzichten en kijkt uit naar de toekomst. Het hoofdstuk beschrijft de drie pilaren (rekenkundige genomics, functionele genomics en moleculaire biologie) die in dit proefschrift zijn gebruikt. Gecombineerd kunnen deze pilaren een veelheid van nieuwe kennis genereren, zowel op het gebied van microbiologie, als op het gebied van bioinformatica. Een aantal bioinformatica voorspellingen zijn in dit proefschrift aan bod gekomen en zijn tevens experimenteel gecontroleerd om de kracht van deze analyses te benadrukken. Tevens kunnen de analyses gebruikt worden om organismen zoals *C. saccharolyticus* meer waterstof te laten produceren. In de toekomst zal het veld van de bioinformatica en de moleculaire (micro)biologie onherroepelijk blijven groeien door onder andere, de ontwikkeling van nieuwe technieken, zoals pyrosequencing, de kostenreductie van het gebruik van deze technieken, de behoefte aan data-integratie en het kwantificeren van de interactie van biomoleculen (systeem biologie).

Dankwoord

Dankwoord

Hierbij wil ik graag de vele mensen bedanken die een bijdrage hebben geleverd aan het tot stand komen van dit proefschrift.

Allereerst natuurlijk John van der Oost, mijn dagelijkse begeleider en tevens promotor. Je sterke uithoudingsvermogen, visie en mogelijkheden die je ziet in het onderzoek zijn van groot belang geweest. Deze eigenschappen hebben mij gesterkt om verder te gaan op het pad dat was uitgezet. Daarnaast ben je zeer goed in de omgang met mensen, wat de sfeer bevordert en het werken in de Bacterial Genetica (BacGen) groep zo aantrekkelijk maakt. Ook mijn tweede promotor Willem de Vos is belangrijk geweest in het opzetten en slagen van dit AIO-project. Jouw begeleiding op soms letterlijk grote afstand wordt zeer gewaardeerd en de snelheid waarop je op nieuwe situaties reageert is bewonderenswaardig. Als derde wil ik de copromotor Servé Kengen bedanken voor de begeleiding van voornamelijk het “Caldi” hoofdstuk. Het was een lange en zware zit, maar heeft wel geleid tot mijn inziens een mooi hoofdstuk, waarbij jouw ervaring en inzet cruciaal waren om Caldi als goede waterstof producent in de etalage te zetten.

Naast de begeleiding gedurende het AIO-schap zijn er vele mensen die een bijdrage hebben geleverd en zonder jullie zou het proefschrift niet mogelijk zijn geweest. First, I want to thank the paranymphs Pawel and Fabian for the lively discussions during the lunch at Unitas, the good atmosphere you generate at BacGen and the assistance during my defense. Verder wil ik de alle mensen van de BacGen groep bedanken. Stan voor de sfeer en de adviezen gedurende het gehele traject. Thijs voor de start van bioinformatica in de groep en je grote enthousiasme. Ans voor je grappen en het draaiende houden van het lab. Jaapie als BacGen DJ zonder uitzendrechten. Corné V. voor je palindromisch oog. Matthijs voor een dosis gezond cynisme. Jasper W. voor het werk aan *Sulfolobus* en je autonomie in het leven. Krisztina voor de uitjes in en naar Rotterdam. Marke, voor de kamersfeer. Bart the Ubuntu King voor je computer hulp, en alle andere medewerkers: Magnus, Ronnie, John glycosaminoglycan is my middle name Raedts, Pierpaolo, Ratnesh, Suzanne, Hao next time when we are in China we will visit you and your family, Corné v.d. K, Colin, Lucy, Katrin, Johan, Judith, Odette, Edze, Ana en Mark & Marco vooral voor de trip to the US of A.

Het proefschrift kon alleen tot stand komen door samenwerking met vele partners waarbij ik, in particular, Marcel and Amy would like to acknowledge for their experimental input for the “Caldi” chapter, and Karin for your annotation and your enthusiasm. Daarnaast hebben Fons, Heleen en Ed belangrijke bijdrages geleverd, as well as many foreign collaborators Karen, Donald, Jason, Emmanuel and Bob. Tevens was er een vruchtbare *Sulfolobus* reconstructie mogelijk door samenwerking die via Bram Snijders tot stand kwam. Moreover, the results of two publications are discussed in this thesis for these fruitful collaborations I want to acknowledge Harry, Tamotsu and Gaël.

Natuurlijk ben ik veel ook dank verschuldigd aan de mensen van microbiologie in zowel de werkgroepen Molecular Ecology, Microbial Physiology, Fungal Genomics als de vaste staf. Met name wil ik Wim bedanken voor al de hardware en software zaken bij microbiologie. Je lijkt altijd tijd te hebben hoe druk je het ook hebt. Tevens wil ik Peter bedanken voor al de bioinformatica vragen die ik had en Hauke voor het helpen bij het maken van fylogenetische bomen. Tot slot heb ik prettig samengewerkt met Petra en Sander in het dagelijks bestuur van de leerstoelgroep.

And of course I don't forget the students from all over the world: Mark, Barzan, Gera, Weilin, Ana, Mark, Barzan. All of you helped me a lot and reflected the diversity of the Wageningen University.

Natuurlijk wil ik mijn ouders bedanken voor hun wekelijkse belangstelling en Nancy en Cécile voor hun interesse en support. Tevens heb ik veel gehad aan mijn vrienden en kennissen voor hun steun en hun belangstelling, waarbij ik speciaal Martijn dankbaar ben voor zijn gastvrijheid gedurende de Rotterdam periode en het brothers in arms gevoel dat we deelden.

Tot slot wil ik graag Ans overweldigend danken. Ans: je steun, je goede adviezen, de vakanties die je voor ons hebt organiseerd en je nooit aflatende opgewektheid hebben mij, de afgelopen jaren, enorm geholpen. De liefde die je geeft is hartverwarmend en voor mij onmisbaar.

Appendix I

Supplementary material

Supplementary material Chapter 4 Identification of a glycolytic regulon in the archaea *Pyrococcus* and *Thermococcus*

Table S4.1. Identified Thermococcales-Glycolytic-Motif (TGM), putative Transcription Factor B-responsive element (BRE) and TATA-box in promoter sequences of *P. furiosus* and *T. kodakaraensis*.

	Pyrococcus furiosus									*Thermococcus kodakaraensis*			
Gene Product	Locus Name	Translation start	Characterization	Start TGM	TGM score[a]	BRE TATA-box	BRE TATA-box score[a]	Tran-scription start site	Intensity Ratio [b]	Locus Name	Start TGM	TGM score[a]	COG ID[c]
Glycolytic genes													
Phospho*enol*pyruvate synthetase	PF0043	(Hutchins *et al.*, 2001)	(Hutchins *et al.*, 2001; Sakuraba *et al.*, 1999)	-38	14.928	-64	9.718	-28	0.18	TK1292	-34	16.919	COG0574
Phosphoglucose isomerase	PF0196[e]	(Verhees *et al.*, 2001)	(Verhees *et al.*, 2001)	-17	14.05	-46	11.327	-11	2.26[d]	TK1111	-67	23.22	COG2140
Enolase	PF0215	(Peak *et al.*, 1994)	(Peak *et al.*, 1994)	-21	15.58	-46	10.772	-	1.32[d]	TK2106	-20	20.735	COG0148
ADP-dependent glucokinase	PF0312	(Kengen *et al.*, 1995)	(Kengen *et al.*, 1995)	-20	17.637	-45	12.024	-10	1.09[d]	TK1110	-20	19.185	COG4809
Glyceraldehyde-3-phosphate: ferredoxin oxidoreductase	PF0464	(Mukund and Adams, 1995)	(Mukund and Adams, 1995)	-35	14.47	-67	14.224	-28	2.54[d]	TK2163	-33	17.942	COG2414
ADP-dependent phosphofructokinase	PF1784		(Tuininga *et al.*, 1999)	-20	15.941	-64	11.895	-27	2.54[d]	TK0376	-20	17.83	COG4809
Triosephosphate isomerase	PF1920[f]			-21	17.359	-49	10.932	-11	2.24[d]	TK2129	-21	17.359	COG0149
Fructose-1,6-bisphosphate aldolase	PF1956[j]		(Siebers *et al.*, 2001)	-20	18.895	-46	13.678	-10	0.33	TK0989	-20	20.484	COG1830
Phosphoglycerate mutase	PF1959		(van der Oost *et al.*, 2002)	-21	17.395	-47	8.32	-	1.75[d]	TK0866	-20	19.333	COG3635
non-glycolytic genes													
α-glucosidase	PF0132	(Badr *et al.*, 1994)	(Badr *et al.*, 1994; Costantino *et al.*, 1990)	-20	16.571	-48	10.178	-	0.71	TK2148	-19	15.561[k]	COG4697
4-α-glucanotransferase	PF0272[gj]	(Laderman *et al.*, 1993b)	(Laderman *et al.*, 1993b; Lee *et al.*, 2006)	-32	17.432	-51	9.901	-	4.7[d]	TK1809	-33	20.484 [k]	COG1449
α-amylase	PF0477[h]			-50	17.598 [k]	-65	15.034	-	-2.45[d]	TK1884	-50	16.646	COG0366
cyclomaltodextrin glucanotransferase	PF0478[i]		(Jorgensen *et al.*, 1997)	-19	18.518	-47	9.774	-	0.89	TK2172	-19	18.162	COG0366

Phosphohexomutase	PF0588		-19	19.171	-46	8.571	-	0.69[d]	TK1108	-29	19.837	COG1109
Fructose-1,6-bisphosphatase	PF0613[j]		-87	15.269	-58	8.973	-	-3.95[d]	TK2164	-64	15.684[k]	COG1980
Hypothetical protein	PF1109		-19	16.849	-48	12.473	-	3.36[d]	-	-	-	COG1572
Maltodextrin binding protein precursor	PF1938	(Evdokimov *et al.*, 2001)	-37	14.954	-65	9.844	-27	2.13[d]	TK1771	-50	18.224	COG2182
predicted transcription regulator, TrmB family	PF0124		-	-		12.703	-	0.4	TK1769[k]	-59	15.935	COG1378
methylmalonyl-CoA decarboxylase, alpha subunit	PF0671		-				-	-1.26[d]	TK1622	-85	17.631	COG4799
ferritin-like protein	PF0742		-				-	-	TK1999	-100	17.883	COG1528
NADP-dependent glyceraldehyde-3-phosphate dehydrogenase (non-phosphorylating)	PF0755		-				-	-0.81	TK0705	-20	17.655	COG1012
Putative Glycoside hydrolase, family 57	PF0870		-				-	0.11	TK1743	-19	19.072	-
Hypothetical protein	PF1025		-				-	2.47[d]	TK1136	-20	15.955	-
probable α-amylase, GH57 family	PF1393		-				-	1.57[d]	TK1436	-19	16.667	COG1543
2-dehydropantoate 2-reductase	PF1396		-				-	-0.89 [d]	TK1968	-66	15.415	COG1893
α-glucan phosphorylase	PF1535		-				-	1.35 [d]	TK1406	-31	16.276	COG0058
pullulanase type II, GH13 family	-		-				-	-	TK0977	-28	20.429	COG0296
phospho-sugar mutase	PF1729		-				-	-0.6	TK1404[k]	-34	19.982	COG1109
predicted thiol protease	-		-				-	-	TK1295	-95	16.928[k]	COG4870
hypothetical protein	-		-				-	-	TK1159	-33	18.389	-

The negative numbers indicate the position of the nucleotides upstream the translation start codon.
[a] Scores are calculated using the search method of TFBS modules (Lenhard and Wasserman, 2002).
[b] Mean intensity ratio ($\log_2$) of maltose-grown cells versus peptide grown cells (Schut *et al.*, 2003)
[c] Cluster of Orthologous Groups (COG) IDs (Tatusov *et al.*, 2003) were assigned with Conserved Domain-search (Marchler-Bauer *et al.*, 2003).
[d] Significantly ($P < 0.01$) up or down-regulated in microarray experiment of cells grown on maltose versus peptides (Schut *et al.*, 2003).
[e] Although TGM and the putative BRE and TATA-box are sited in the coding sequence of PF0195, it is not one transcription unit (this study) and therefore retained as positive hit.
[f] Characterized in *Pyrococcus woesei* (Kohlhoff *et al.*, 1996)
[g] Translation start site correction of 21 nucleotides downstream, compared to genome sequence annotation (Laderman *et al.*, 1993a).
[h] Translation start site correction of 39 nucleotides downstream, compared to genome sequence annotation (Jorgensen *et al.*, 1997).
[i] Translation start site correction of 97 nucleotides and frame shift at nucleotide 53 upstream, compared to genome sequence annotation.
[j] These genes are divergently oriented with another gene, which is supposedly not under control of the same transcriptional regulator.
[k] Score is based on complementary strand.

Table S4.2 Identified Thermococcales-Glycolytic-Motif (TGM) in the order *Thermococcales* in promoter sequences of genes not reported as part of sequenced genomes

Species	Accession Number	Gene Name	Gene Product	Similar to protein in P. furiosus	Similar to protein in T. kodakaraensis	Start TGM	TGM score[a]
Thermococcus litoralis	AB050016	*adp-pfk*	ADP-dependent phosphofructokinase	PF1784	TK0376	-19	15.756
Thermococcus litoralis	D88253	*jgt*	4-α-glucanotransferase	PF0272	TK1809	-38	19.837
Thermococcus sp. B1001	AB025721	*cgtA*	cyclodextrin glucanotransferase	PF0478	TK2172	-54	16.282
Thermococcus sp. B1001	AB034969	*cgtC*	cyclomaltodextrin binding protein	PF1938	TK1771	-51	15.288
Thermococcus sp. Rt3	AF017454	*amy*	amylase	PF0477	TK1884	-64	16.276
Thermococcus zilligii	AY005811	*pfk*	ADP-dependent phosphofructokinase	PF1784	TK0376	-20	18.482
Thermococcus aggregans	AJ251532	*pulhA*	pullulan hydrolase type III	-	TK0977	-22	18.162
Thermococcus hydrothermalis	AF068255	*amy*	α-amylase	PF0477	TK1884	-44	16.276
Pyrococcus furiosus	X80819	*ppsA*	pyruvate,water dikinase	PF0043	TK1292	-38	18.38
Pyrococcus furiosus	L22346	*amyA*	α-amylase	PF0272	TK1809	-32	18.989
Pyrococcus sp.	D83793	*apkA*	α-amylase	PF0477	TK1884	-50	15.684
Pyrococcus woesei	AF177906	*amyA*	α-amylase	PF0477	TK1884	-50	16.783
Pyrococcus woesei	AF240464	*pow*	α-amylase	PF0477	TK1884	-50	16.783

The negative numbers indicate the position of the nucleotides upstream the translation start codon.
[a] Scores are calculated using the search method of TFBS modules (Lenhard and Wasserman, 2002).

Supplementary material Chapter 5 Hydrogenomics of the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus*

Table S5.1 Taxonomic distribution of species with more than 100 best blast hits, based on Blastp against the KEGG database.

Species	**no**
Thermoanaerobacter tengcongensis	549
Clostridium thermocellum	512
Carboxydothermus hydrogenoformans	145
Clostridium acetobutylicum	140

Table S5.2 Genome properties of organisms used for comparative genomics based on IMG data (Markowitz *et al.*, 2006).

Organism	**Chromosome**	**Size (kbp)**	**Proteins**[a]	**GC-content (%)**
Caldicellulosiruptor saccharolyticus DSM 8903	1	2,970	2,679	32.3
Clostridium thermocellum ATCC 27405	1	3,843	3,191	39.0
Thermoanaerobacter tengcongensis MB4	1	2,689	2,588	37.6
Thermotoga maritima MSB8	1	1,861	1,858	46.3
Pyrococcus furiosus DSM 3638	1	1,908	1,983	40.8

[a]Number of predicted proteins without pseudogenes.

Table S5.3 Carbohydrate-active enzymes encoded by the genome of *Caldicellulosiruptor saccharolyticus*

Locus id	Protein name	EC[a]	Family[b]	CBM[b]
Csac_1089	β-glucosidase	3.2.1.21	GH1	
Csac_0129	β-mannosidase	3.2.1.25	GH2	
Csac_0362	β-galactosidase	3.2.1.23	GH2	
Csac_2686	putative β-galactosidase/β-glucuronidase		GH2	
Csac_2734	β-galactosidase	3.2.1.23	GH2	
Csac_0586	β-glucosidase/xylan 1,4-β-xylosidase	3.2.1.21/3.2.1.37	GH3	
Csac_1102	β-glucosidase	3.2.1.21	GH3	
Csac_1562	α-galactosidase (melibiase)	3.2.1.22	GH4	
Csac_2748	putative α-galactosidase		GH4	
Csac_0137	cellulase	3.2.1.4	GH5	
Csac_0678	cellulase precursor	3.2.1.4	GH5	
Csac_1080*	mannan endo-1,4-β-mannosidase	3.2.1.78	GH5	
Csac_2528	glycoside hydrolase, family 5		GH5	
Csac_1078	cellulase/cellulose 1,4-β-cellobiosidase precursor	3.2.1.4/3.2.1.91	GH5/10	CBM3
Csac_1077	mannan endo-1,4-β-mannosidase/cellulase precursor	3.2.1.78/3.2.1.4	GH5/44	CBM3/3
Csac_1079	cellulase/mannan endo-1,4-β-mannosidase precursor	3.2.1.4	GH9	CBM3/3
Csac_1076	cellulase precursor	3.2.1.4	GH9/48	CBM3/3
Csac_0204	putative endo-1,4-β-xylanase		GH10	
Csac_2405	endo-1,4-β-xylanase precursor	3.2.1.8	GH10	
Csac_2408	endo-1,4-β-xylanase precursor	3.2.1.8	GH10	
Csac_0696	endo-1,4-β-xylanase precursor	3.2.1.8	GH10	CBM22/22
Csac_2410	endo-1,4-β-xylanase precursor	3.2.1.8	GH10	CBM22/22
Csac_0203	4-α-glucanotransferase	2.4.1.25	GH13	
Csac_0408	α-amylase precursor	3.2.1.1	GH13	
Csac_0426	α-amylase	3.2.1.1	GH13	
Csac_2428	oligo-1,6-glucosidase	3.2.1.10	GH13	
Csac_0689	pullulanase	3.2.1.41	GH13	CBM20/41
Csac_0671	pullulanase	3.2.1.41	GH13	CBM48
Csac_0784	1,4-α-glucan branching enzyme	2.4.1.18	GH13	CBM48
Csac_0130	putative glucan 1,4-α-glucosidase		GH15	
Csac_2548*	putative endo-1,3(4)-β-glucanase		GH16	CBM22/22/22/22/22
Csac_2539	putative β-N-acetylhexosaminidase		GH20	
Csac_1720	protein containing lytic transglycosylase SLT domain		GH23	
Csac_1986	protein containing lytic transglycosylase SLT domain		GH23	
Csac_0663	mannan endo-1,4-β-mannosidase	3.2.1.78	GH26	
Csac_0361	galacturan 1,4-α-galacturonidase	3.2.1.67	GH28	
Csac_0664	putative galacturan 1,4-α-galacturonidase		GH28	
Csac_1340	α-L-fucosidase	3.2.1.51	GH29	
Csac_2513	glycoside hydrolase, family 30		GH30	
Csac_1354	α-xylosidase	3.2.1.-	GH31	
Csac_1118	α-galactosidase	3.2.1.22	GH36	
Csac_2404	xylan 1,4-β-xylosidase	3.2.1.37	GH39	
Csac_2409	xylan 1,4-β-xylosidase	3.2.1.37	GH39	
Csac_1018	β-galactosidase	3.2.1.23	GH42	
Csac_0359	glycoside hydrolase, family 43		GH43	

Csac_1560	arabinan endo-1,5-α-L-arabinosidase	3.2.1.99	GH43	
Csac_0437	glycoside hydrolase, family 43		GH43	
Csac_2411	xylan 1,4-β-xylosidase/α-N-arabinofuranosidase precursor	3.2.1.37/3.2.1.55	GH43/43	CBM22/6
Csac_1561	α-N-arabinofuranosidase	3.2.1.55	GH51	
Csac_0439	kojibiose phosphorylase	2.4.1.230	GH65	
Csac_0444	putative trehalose/maltose hydrolase (possible phosphorylases)		GH65	
Csac_2689	α-glucuronidase	3.2.1.139	GH67	
Csac_1085	glycoside hydrolase, family 74 with carbohydrate-binding module protein precursor		GH74	CBM3
Csac_1105/ Csac_1107	α-L-rhamnosidase N-terminal domain/C-terminal domain protein	3.2.1.40	GH78	
Csac_2730	unsaturated glucuronyl hydrolase		GH88	
Csac_1090	putative cellobiose phosphorylase		GH94	
Csac_1091	cellobiose phosphorylase	2.4.1.20	GH94	
Csac_0360	unsaturated rhamnogalacturonyl hydrolase	3.2.1.-	GH105	
Csac_0206	putative acetylesterase		IPR005181	
Csac_0258	glycosidase, PH1107-related		IPR007184	
Csac_0259	glycosidase, PH1107-related		IPR007184	
Csac_0296	glycosidase, PH1107-related		IPR007184	
Csac_0762	glycosidase, PH1107-related		IPR007184	
Csac_0853	glycosidase, PH1107-related		IPR007184	
Csac_2527	glycosidase, PH1107-related		IPR007184	
Csac_2519	putative glycoside hydrolase precursor with carbohydrate-binding module		IPR013781	CBM28
Csac_0268	glycosyltransferase, family 2		GT2	
Csac_1057	glycosyltransferase, family 2		GT2	
Csac_1681	glycosyltransferase, family 2		GT2	
Csac_1877	glycosyltransferase, family 2		GT2	
Csac_2350	glycosyltransferase, family 2		GT2	
Csac_2426	glycosyltransferase, family 2		GT2	
Csac_2567	glycosyltransferase, family 2		GT2	
Csac_2631	glycosyltransferase, family 2		GT2	
Csac_1349	glycosyltransferase WecB/TagA/CpsF family and Polysaccharide pyruvyl transferase domain protein		GT26	
Csac_0925	undecaprenyldiphospho-muramoylpentapeptide β-N-acetylglucosaminyltransferase	2.4.1.227	GT28	
Csac_2337	1,2-diacylglycerol 3-glucosyltransferase homolog		GT28	
Csac_0429	α-1,4-glucan phosphorylase	2.4.1.1	GT35	
Csac_0780	glycogen phosphorylase	2.4.1.1	GT35	
Csac_1081	glycosyl transferase, family 39		GT39	
Csac_0134	glycosyltransferase, family 4		GT4	
Csac_0194	glycosyltransferase, family 4		GT4	
Csac_1092	glycosyltransferase, family 4		GT4	
Csac_1346	glycosyltransferase, family 4		GT4	
Csac_1682	glycosyltransferase, family 4		GT4	
Csac_1729	glycosyltransferase, family 4		GT4	
Csac_1745	glycosyltransferase, family 4		GT4	
Csac_1808	glycosyltransferase, family 4		GT4	

Csac_2361	glycosyltransferase, family 4		GT4
Csac_2568	glycosyltransferase, family 4		GT4
Csac_2569	glycosyltransferase, family 4		GT4
Csac_2570	glycosyltransferase, family 4		GT4
Csac_2572	glycosyltransferase, family 4		GT4
Csac_0781	glycogen synthase	2.4.1.21	GT5
Csac_0371	membrane carboxypeptidase (penicillin-binding protein)		GT51
Csac_0490	membrane carboxypeptidase (penicillin-binding protein)		GT51
Csac_2407	acetylesterase	3.1.1.6	CE1
Csac_0205	polysaccharide deacetylase family protein		CE4
Csac_0719	polysaccharide deacetylase family protein		CE4
Csac_2009	polysaccharide deacetylase family protein		CE4
Csac_2371	polysaccharide deacetylase family protein		CE4
Csac_2436	acetylxylan esterase	3.1.1.72	CE7
Csac_0213	putative amidohydrolase		CE9
Csac_2538	N-acetylglucosamine-6-phosphate deacetylase	3.5.1.25	CE9

Truncated genes are indicated by an asterisk and may not be functional. [a]EC enzyme commission number; [b]Proteins are grouped by Carbohydrate-Active enzymes (CAZy; http://www.cazy.org (Coutinho and Henrissat, 1999)) classification: Glycoside Hydrolase (GH), Glycosyltransferase (GT), Carbohydrate Esterases (CE), Carbohydrate-Binding Modules (CBM), no Polysaccharide lyases were detected. The list was extended with InterPro (IPR) GH-families (Mulder *et al.*, 2007), which are not in the CAZy database.

Table S5.4 Relative expression of ORFs in cells grown on different carbon sources. Only ORFs whose expression is dramatically up- or downregulated are shown ($\log_2$-value > 2 or < -2, respectively).

Glucose versus rhamnose

ORF	Protein name	Function	Intensity ratio (log2)	Change in expression (fold)
Csac_2506	xylose/glucose ABC transporter, periplasmic component	ABC transporter	2.45	5.47
Csac_2505	xylose/glucose ABC transporter, ATPase component	ABC transporter	2.12	4.33
Csac_0243	conserved hypothetical protein		2.50	5.64
Csac_0346	ATP-binding region, ATPase-like protein		2.01	4.04
Csac_0431	putative maltodextrin ABC transport system, periplasmic component	maltosedextrin utilization	2.45	5.47
Csac_0622	iron-containing alcohol dehydrogenase		3.56	11.80
Csac_0880	ATP-dependent Clp protease, ATP-binding subunit ClpX		2.07	4.19
Csac_1189	fructose-1,6-bisphosphate aldolase, class II	glycolysis	2.37	5.18
Csac_1461	pyruvate:ferredoxin oxidoreductase subunit beta	glycolysis	2.10	4.28
Csac_1606	Acyl carrier protein (ACP)		2.03	4.07
Csac_1628	Conserved hypothetical protein		3.30	9.86
Csac_1823	Predicted metal-binding, possibly nucleic acid-binding protein		2.05	4.13
Csac_1824	ribosomal protein L32	ribosome complex	2.15	4.44
Csac_1846	hypothetical protein		2.20	4.61
Csac_1864	NADH-dependent Fe-only hydrogenase subunit A	hydrogenase (NADH)	3.23	9.35
Csac_1955	pyruvate, phosphate dikinase	glycolysis	2.83	7.13
Csac_1990	Rubredoxin		2.01	4.04

Csac_1991	phosphoribosylamine-glycine ligase	purine synthesis	2.26	4.78
Csac_1997	phosphoribosylformylglycinamidine synthase I	purine synthesis	2.06	4.16
Csac_2000	putative purine permease	purine salvage	3.91	15.07
Csac_2039	desulfoferrodoxin		2.06	4.18
Csac_2040	acetate kinase	glycolysis	2.29	4.90
Csac_2044	Hfq protein	RNA binding	2.35	5.09
Csac_2204	50S ribosomal protein L7/L12	ribosome complex	2.80	6.98
Csac_2205	Ribosomal protein L10 homolog		2.13	4.39
Csac_2450	conserved secreted protein (prefoldin like domain)		2.26	4.78
Csac_2488	Putative carbamoyl-phosphate synthase large chain		2.00	4.01
Csac_0375	Circadian clock protein kinase kaiC (EC 2.7.1.37)., putative		-2.10	4.28
Csac_0407	putative lactaldehyde reductase		-3.19	9.10
Csac_0476	DNA binding protein, putative transcriptional regulator		-2.05	4.13
Csac_0792	3,4-dihydroxy-2-butanone-4-phosphate synthase/ GTP cyclohydrolase II	Flavin biosynthesis	-2.49	5.63
Csac_0793	riboflavin synthase, alpha subunit	Flavin biosynthesis	-2.05	4.15
Csac_0865	rhamnulose-1-phosphate aldolase	rhamnose pathway	-3.94	15.39
Csac_0866	Class II Aldolase and Adducin N-terminal domain protein	rhamnose pathway	-4.66	25.28
Csac_0868	sorbitol-6-phosphate 2-dehydrogenase		-2.71	6.55
Csac_0870	Lipoate-protein ligase B	keto-acid dehydrogenase	-2.87	7.30
Csac_0871	Lipoic acid synthetase	keto-acid dehydrogenase	-4.37	20.71
Csac_0872	Dihydrolipoamide S-acetyltransferase (E2 component)	keto-acid dehydrogenase	-5.17	36.00
Csac_0873	Dihydrolipoamide dehydrogenase (E3 component)	keto-acid dehydrogenase	-4.12	17.34
Csac_0874	Acetoin/Pyruvate/2-oxoglutarate dehydrogenase complex (E1 component)	keto-acid dehydrogenase	-3.01	8.03
Csac_0875	transcriptional regulator, DeoR family	keto-acid dehydrogenase	-4.56	23.55
Csac_0876	L-rhamnose isomerase	rhamnose pathway	-2.57	5.96
Csac_1146	2-isopropylmalate synthase/homocitrate synthase family protein		-2.94	7.67
Csac_1164	Methionine synthase (vitamin-B12 dependent),		-2.15	4.44
Csac_1224	Formate--tetrahydrofolate ligase (Formyltetrahydrofolate synthetase)		-2.00	4.00
Csac_1633	Putative binding-protein-dependent transport systems inner membrane comp.		-2.02	4.07
Csac_1635	Molybdopterin synthase sulfurylase		-2.25	4.75
Csac_2698	phospho-2-dehydro-3-deoxyheptonate aldolase	amino acid biosynthesis	-2.76	6.77
Csac_2699	Prephenate dehydrogenase	amino acid biosynthesis	-2.20	4.58

Glucose versus xylose				
ORF	Protein name	Function	Intensity ratio (log2)	Change in expression (fold)
Csac_0431	putative maltodextrin ABC transport system, periplasmic component	maltodextrin utilization	2.05	4.13
Csac_1627	Hypothetical protein	purine biosynthesis ?	3.33	10.07
Csac_1628	Conserved hypothetical protein	purine biosynthesis ?	4.10	17.10
Csac_1992	phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase	purine biosynthesis	2.60	6.08
Csac_1993	phosphoribosylglycinamide formyltransferase	purine biosynthesis	2.39	5.25
Csac_1994	phosphoribosylformylglycinamidine cyclo-ligase	purine biosynthesis	2.13	4.37
Csac_1995	amidophosphoribosyltransferase	purine biosynthesis	2.90	7.45
Csac_1996	phosphoribosylformylglycinamidine synthase II	purine biosynthesis	2.07	4.20
Csac_1997	phosphoribosylformylglycinamidine synthase I	purine biosynthesis	2.16	4.48
Csac_2000	putative purine permease	purine salvage	3.32	10.02
Csac_0240	ribose/xylose/arabinose/galactoside ABC-type transport systems, ATPase component	ABC transporter	-4.73	0.04
Csac_0241	ribose/xylose/arabinose/galactoside ABC-type transport systems, permease component	ABC transporter	-4.31	0.05
Csac_0242	ribose/xylose/arabinose/galactoside ABC-type transport systems, periplasmic component	ABC transporter	-4.68	0.04
Csac_0692	ABC-type sugar transport system, periplasmic component	xylan/xylose utilization	-4.86	0.03
Csac_0693	ABC-type sugar transport systems, permease component	xylan/xylose utilization	-2.69	0.16
Csac_0694	ABC-type sugar transport system, permease component	xylan/xylose utilization	-2.06	0.24
Csac_0695	putative xylose repressor	xylan/xylose utilization	-2.50	0.18
Csac_0696	endo-1,4-β-xylanase precursor	xylan/xylose utilization	-2.63	0.16
Csac_1154	putative xylose isomerase	xylan/xylose utilization	-2.91	0.13

Xylose versus mixture				
ORF	Protein name	Function	Intensity ratio (log2)	Change in expression (fold)
Csac_0792	3,4-dihydroxy-2-butanone-4-phosphate synthase/ GTP cyclohydrolase II	flavin biosynthesis	-2.82	7.04
Csac_0793	riboflavin synthase, alpha subunit	flavin biosynthesis	-2.52	5.74
Csac_0794	riboflavin biosynthesis protein RibD	flavin biosynthesis	-2.09	4.27
Csac_1499	FMN-dependent NADH-azoreductase		-2.04	4.10

Glucose versus mixture

ORF	Protein name	Function	Intensity ratio (log2)	Change in expression (fold)
Csac_1627	Hypothetical protein	purine biosynthesis ?	2.91	7.50
Csac_1628	Conserved hypothetical protein	purine biosynthesis ?	4.13	17.46
Csac_1992	phosphoribosylaminoimidazolecarboxamide formyltransferase/IMP cyclohydrolase	purine biosynthesis	2.52	5.72
Csac_1993	phosphoribosylglycinamide formyltransferase	purine biosynthesis	2.23	4.69
Csac_1995	amidophosphoribosyltransferase	purine biosynthesis	2.51	5.70
Csac_1996	phosphoribosylformylglycinamidine synthase II	purine biosynthesis	2.11	4.32
Csac_1997	phosphoribosylformylglycinamidine synthase I	purine biosynthesis	2.19	4.57
Csac_2000	putative purine permease	purine salvage	2.01	4.02
Csac_0240	ribose/xylose/arabinose/galactoside ABC-type transport systems, ATPase component	ABC transporter	-3.06	8.35
Csac_0241	ribose/xylose/arabinose/galactoside ABC-type transport systems, permease component	ABC transporter	-3.89	14.81
Csac_0242	ribose/xylose/arabinose/galactoside ABC-type transport systems, periplasmic component	ABC transporter	-4.00	16.01
Csac_0692	ABC-type sugar transport system, periplasmic component	xylan/xylose utilization	-5.26	38.34
Csac_0693	ABC-type sugar transport systems, permease component	xylan/xylose utilization	-2.06	4.18
Csac_0694	ABC-type sugar transport system, permease component	xylan/xylose utilization	-2.14	4.39
Csac_0695	putative xylose repressor	xylan/xylose utilization	-2.19	4.55
Csac_0696	endo-1,4-β-xylanase precursor	xylan/xylose utilization	-2.72	6.59
Csac_0792	3,4-dihydroxy-2-butanone-4-phosphate synthase/ GTP cyclohydrolase II	flavin biosynthesis	-3.31	9.94
Csac_0793	riboflavin synthase, alpha subunit	flavin biosynthesis	-2.67	6.36
Csac_0794	riboflavin biosynthesis protein RibD	flavin biosynthesis	-2.17	4.49
Csac_0870	Lipoate-protein ligase B	keto-acid dehydrogenase	-2.01	4.04
Csac_0871	Lipoic acid synthetase	keto-acid dehydrogenase	-2.84	7.14
Csac_0873	Dihydrolipoamide dehydrogenase (E3 component)	keto-acid dehydrogenase	-1.97	3.93
Csac_0875	transcriptional regulator, DeoR family	rhamnose pathway	-2.30	4.94
Csac_1154	putative xylose isomerase	xylan/xylose utilization	-2.55	5.85

Mixture versus rhamnose				
ORF	**Protein name**	**Function**	**Intensity ratio (log2)**	**Change in expression (fold)**
Csac_0053	Transposase, Mutator family		2.20	4.59
Csac_0240	ribose/xylose/arabinose/galactoside ABC-type transport systems, ATPase component	ABC transporter	3.74	13.38
Csac_0241	ribose/xylose/arabinose/galactoside ABC-type transport systems, permease component	ABC transporter	4.73	26.59
Csac_0242	ribose/xylose/arabinose/galactoside ABC-type transport systems, periplasmic component	ABC transporter	4.66	25.30
Csac_0243	conserved hypothetical protein		2.13	4.38
Csac_0346	ATP-binding region, ATPase-like protein		2.43	5.38
Csac_0622	iron-containing alcohol dehydrogenase		2.74	6.69
Csac_0692	ABC-type sugar transport system, periplasmic component	xylan/xylose utilization	5.72	52.86
Csac_0693	ABC-type sugar transport systems, permease component	xylan/xylose utilization	2.31	4.95
Csac_0694	ABC-type sugar transport system, permease component	xylan/xylose utilization	2.01	4.03
Csac_0695	putative xylose repressor	xylan/xylose utilization	2.14	4.41
Csac_0696	endo-1,4-β-xylanase precursor	xylan/xylose utilization	2.94	7.66
Csac_1211	glutamine synthetase, type 1		2.52	5.73
Csac_1302	conserved hypothetical protein		2.18	4.52
Csac_1305	Protein of unknown function (DUF1015)		2.35	5.11
Csac_1324	Ribosome-associated protein Y	ribosome complex	2.05	4.14
Csac_1335	30S ribosomal protein S6	ribosome complex	2.17	4.51
Csac_1520	ribosomal protein L31	ribosome complex	2.45	5.47
Csac_1606	Acyl carrier protein (ACP)		2.09	4.25
Csac_1846	hypothetical protein		2.35	5.09
Csac_1864	NADH-dependent Fe-only hydrogenase subunit A	hydrogenase (NADH)	3.90	14.97
Csac_1953	glyceraldehyde-3-phosphate dehydrogenase	glycolysis	2.02	4.07
Csac_1955	pyruvate, phosphate dikinase	glycolysis	2.97	7.83
Csac_2039	desulfoferrodoxin		2.02	4.05
Csac_2040	acetate kinase	glycolysis	2.04	4.12
Csac_2044	Hfq protein	RNA binding	2.83	7.13
Csac_2109	CoA binding domain		2.33	5.01
Csac_2204	50S ribosomal protein L7/L12	ribosome complex	3.21	9.26
Csac_2205	Ribosomal protein L10 homolog	ribosome complex	2.59	6.02
Csac_2450	conserved secreted protein (prefoldin like domain)		2.25	4.75
Csac_2488	Putative carbamoyl-phosphate synthase large chain		2.28	4.86
Csac_2504	xylose/glucose ABC transporter, permease component	ABC transporter	2.26	4.80
Csac_2505	xylose/glucose ABC transporter, ATPase component	ABC transporter	2.28	4.85
Csac_2506	xylose/glucose ABC transporter, periplasmic component	ABC transporter	2.60	6.05

Csac_0407	putative lactaldehyde reductase		-1.96	3.89
Csac_0865	rhamnulose-1-phosphate aldolase	rhamnose pathway	-3.95	15.42
Csac_0866	Class II Aldolase and Adducin N-terminal domain protein	rhamnose pathway	-4.14	17.69
Csac_0872	Dihydrolipoamide S-acetyltransferase (E2 component)	keto-acid dehydrogenase	-3.46	10.99
Csac_0873	Dihydrolipoamide dehydrogenase (E3 component)	keto-acid dehydrogenase	-2.14	4.41
Csac_0875	transcriptional regulator, DeoR family	keto-acid dehydrogenase	-2.25	4.77

Xylose versus rhamnose

ORF	Protein name	Function	Intensity ratio (log2)	Change in expression (fold)
Csac_0053	Transposase, Mutator family		2.33	5.02
Csac_0240	ribose/xylose/arabinose/galactoside ABC-type transport systems, ATPase component	ABC transporter	5.42	42.69
Csac_0241	ribose/xylose/arabinose/galactoside ABC-type transport systems, permease component	ABC transporter	5.16	35.67
Csac_0242	ribose/xylose/arabinose/galactoside ABC-type transport systems, periplasmic component	ABC transporter	5.34	40.62
Csac_0243	conserved hypothetical protein		3.18	9.09
Csac_0622	iron-containing alcohol dehydrogenase		2.71	6.55
Csac_0650	hypothetical protein		2.02	4.06
Csac_0692	ABC-type sugar transport system, periplasmic component	xylan/xylose utilization	5.32	40.06
Csac_0693	ABC-type sugar transport systems, permease component	xylan/xylose utilization	2.93	7.64
Csac_0695	putative xylose repressor	xylan/xylose utilization	2.46	5.49
Csac_0696	endo-1,4-β-xylanase precursor	xylan/xylose utilization	2.84	7.17
Csac_0798	xylulokinase	xylan/xylose utilization	2.10	4.28
Csac_0930	cell division protein FtsZ		1.96	3.90
Csac_1154	putative xylose isomerase	xylan/xylose utilization	2.09	4.24
Csac_1211	glutamine synthetase, type I		2.43	5.38
Csac_1335	30S ribosomal protein S6	ribosome complex	2.50	5.66
Csac_1460	pyruvate:ferredoxin oxidoreductase subunit alpha		2.17	4.51
Csac_1461	pyruvate:ferredoxin oxidoreductase subunit beta	glycolysis	2.07	4.19
Csac_1520	ribosomal protein L31	ribosome complex	2.30	4.91
Csac_1580	reverse gyrase		2.04	4.12
Csac_1824	ribosomal protein L32	ribosome complex	2.24	4.73
Csac_1864	NADH-dependent Fe-only hydrogenase subunit A	hydrogenase (NADH)	3.17	9.01
Csac_1955	pyruvate, phosphate dikinase	glycolysis	2.68	6.41

Csac_2040	acetate kinase	glycolysis	2.30	4.93
Csac_2044	Hfq protein	RNA binding	3.13	8.78
Csac_2204	50S ribosomal protein L7/L12	ribosome complex	2.79	6.94
Csac_2205	Ribosomal protein L10 homolog	ribosome complex	2.45	5.45
Csac_2450	conserved secreted protein (prefoldin like domain)		2.47	5.55
Csac_2488	Putative carbamoyl-phosphate synthase large chain		2.68	6.40
Csac_2505	xylose/glucose ABC transporter, ATPase component	ABC transporter	2.27	4.81
Csac_2506	xylose/glucose ABC transporter, periplasmic component	ABC transporter	2.65	6.26
Csac_0407	putative lactaldehyde reductase		-2.51	5.70
Csac_0432	Hypothetical protein		-2.06	4.16
Csac_0792	3,4-dihydroxy-2-butanone-4-phosphate synthase/ GTP cyclohydrolase II	riboflavin biosynthesis	-2.00	3.99
Csac_0865	rhamnulose-1-phosphate aldolase	rhamnose pathway	-3.74	13.39
Csac_0866	Class II Aldolase and Adducin N-terminal domain protein	rhamnose pathway	-4.77	27.31
Csac_0870	Lipoate-protein ligase B	keto-acid dehydrogenase	-2.15	4.45
Csac_0871	Lipoic acid synthetase	keto-acid dehydrogenase	-2.61	6.13
Csac_0872	Dihydrolipoamide S-acetyltransferase (E2 component)	keto-acid dehydrogenase	-4.19	18.23
Csac_0873	Dihydrolipoamide dehydrogenase (E3 component)	keto-acid dehydrogenase	-2.86	7.28
Csac_0874	Acetoin/Pyruvate/2-oxoglutarate dehydrogenase complex (E1 component)	keto-acid dehydrogenase	-2.08	4.23
Csac_0875	transcriptional regulator, DeoR family	rhamnose pathway	-3.27	9.63
Csac_0876	L-rhamnose isomerase	rhamnose pathway	-2.15	4.44
Csac_1030	Oligopeptide transport ATP-binding protein oppD		-2.12	4.34
Csac_1146	2-isopropylmalate synthase/homocitrate synthase family protein		-2.33	5.03
Csac_1224	Formate--tetrahydrofolate ligase (Formyltetrahydrofolate synthetase)		-2.35	5.09
Csac_1635	Molybdopterin synthase sulfurylase		-2.00	4.01
Csac_2590	LemA family protein		-2.41	5.30
Csac_2698	phospho-2-dehydro-3-deoxyheptonate aldolase	amino acid biosynthesis	-2.28	4.85

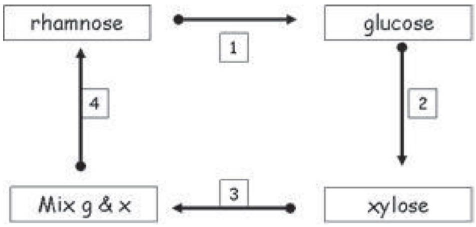


Figure S5.1 The experimental 4-loop design of *C. saccharolyticus* grown on L-rhamnose, glucose, xylose and a mix of xylose and glucose.

References

Badr, H. R., Sims, K. A., and Adams, M. W. W. (1994). Purification and Characterization of Sucrose α-Glucohydrolase (Invertase) from the Hyperthermophilic Archaeon *Pyrococcus furiosus*. Systematic and Applied Microbiology *17*, 1-6.

Costantino, H. R., Brown, S. H., and Kelly, R. M. (1990). Purification and characterization of an α-glucosidase from a hyperthermophilic archaebacterium, *Pyrococcus furiosus*, exhibiting a temperature optimum of 105 to 115 degrees C. J Bacteriol *172*, 3654-3660.

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

Evdokimov, A. G., Anderson, D. E., Routzahn, K. M., and Waugh, D. S. (2001). Structural basis for oligosaccharide recognition by *Pyrococcus furiosus* maltodextrin-binding protein. J Mol Biol *305*, 891-904.

Hutchins, A. M., Holden, J. F., and Adams, M. W. (2001). Phosphoenolpyruvate synthetase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol *183*, 709-715.

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

Kengen, S. W., Tuininga, J. E., de Bok, F. A., Stams, A. J., and de Vos, W. M. (1995). Purification and characterization of a novel ADP-dependent glucokinase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Biol Chem *270*, 30453-30457.

Kohlhoff, M., Dahm, A., and Hensel, R. (1996). Tetrameric triosephosphate isomerase from hyperthermophilic Archaea. FEBS Lett *383*, 245-250.

Laderman, K. A., Asada, K., Uemori, T., Mukai, H., Taguchi, Y., Kato, I., and Anfinsen, C. B. (1993a). α-amylase from the hyperthermophilic archaebacterium *Pyrococcus furiosus*. Cloning and sequencing of the gene and expression in *Escherichia coli*. J Biol Chem *268*, 24402-24407.

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

Lee, H. S., Shockley, K. R., Schut, G. J., Conners, S. B., Montero, C. I., Johnson, M. R., Chou, C. J., Bridger, S. L., Wigner, N., Brehm, S. D*., et al.* (2006). Transcriptional and biochemical analysis of starch metabolism in the hyperthermophilic archaeon *Pyrococcus furiosus*. J Bacteriol *188*, 2115-2125.

Lenhard, B., and Wasserman, W. W. (2002). TFBS: Computational framework for transcription factor binding site analysis. Bioinformatics *18*, 1135-1136.

Marchler-Bauer, A., Anderson, J. B., DeWeese-Scott, C., Fedorova, N. D., Geer, L. Y., He, S., Hurwitz, D. I., Jackson, J. D., Jacobs, A. R., Lanczycki, C. J*., et al.* (2003). CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res *31*, 383-387.

Markowitz, V. M., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I*., et al.* (2006). The integrated microbial genomes (IMG) system. Nucleic Acids Res *34*, D344-348.

Mukund, S., and Adams, M. W. (1995). Glyceraldehyde-3-phosphate ferredoxin oxidoreductase, a novel tungsten-containing enzyme with a potential glycolytic role in the hyperthermophilic archaeon *Pyrococcus furiosus*. J Biol Chem *270*, 8389-8392.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Buillard, V., Cerutti, L., Copley, R*., et al.* (2007). New developments in the InterPro database. Nucleic Acids Res *35*, D224-228.

Peak, M. J., Peak, J. G., Stevens, F. J., Blamey, J., Mai, X., Zhou, Z. H., and Adams, M. W. (1994). The hyperthermophilic glycolytic enzyme enolase in the archaeon, *Pyrococcus furiosus*: comparison with mesophilic enolases. Arch Biochem Biophys *313*, 280-286.

Sakuraba, H., Utsumi, E., Kujo, C., and Ohshima, T. (1999). An AMP-dependent (ATP-forming) kinase in the hyperthermophilic archaeon *Pyrococcus furiosus*: characterization and novel physiological role. Arch Biochem Biophys *364*, 125-128.

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

Siebers, B., Brinkmann, H., Dorr, C., Tjaden, B., Lilie, H., van der Oost, J., and Verhees, C. H. (2001). Archaeal fructose-1,6-bisphosphate aldolases constitute a new family of archaeal type class I aldolase. J Biol Chem *276*, 28710-28718.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N.*, et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics *4*, 41.

Tuininga, J. E., Verhees, C. H., van der Oost, J., Kengen, S. W., Stams, A. J., and de Vos, W. M. (1999). Molecular and biochemical characterization of the ADP-dependent phosphofructokinase from the hyperthermophilic archaeon *Pyrococcus furiosus*. J Biol Chem *274*, 21023-21028.

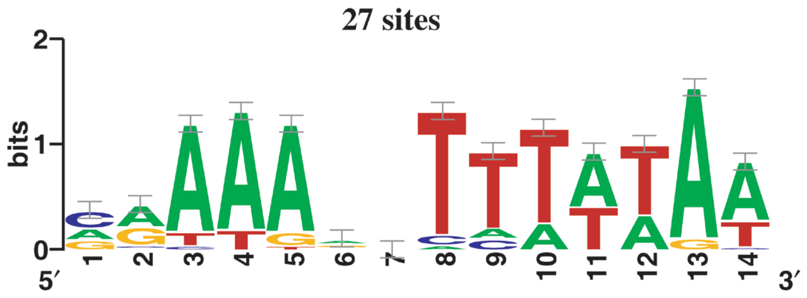
van der Oost, J., Huynen, M. A., and Verhees, C. H. (2002). Molecular characterization of phosphoglycerate mutase in archaea. FEMS Microbiol Lett *212*, 111-120.

Verhees, C. H., Huynen, M. A., Ward, D. E., Schiltz, E., de Vos, W. M., and van der Oost, J. (2001). The phosphoglucose isomerase from the hyperthermophilic archaeon *Pyrococcus furiosus* is a unique glycolytic enzyme that belongs to the cupin superfamily. J Biol Chem *276*, 40926-40932.

Appendix II

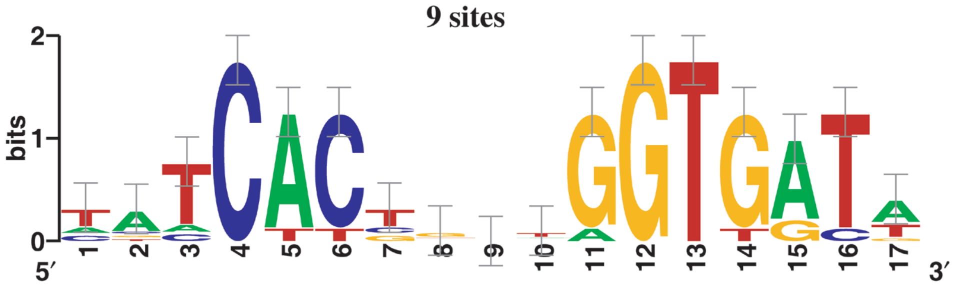
Color figures

A



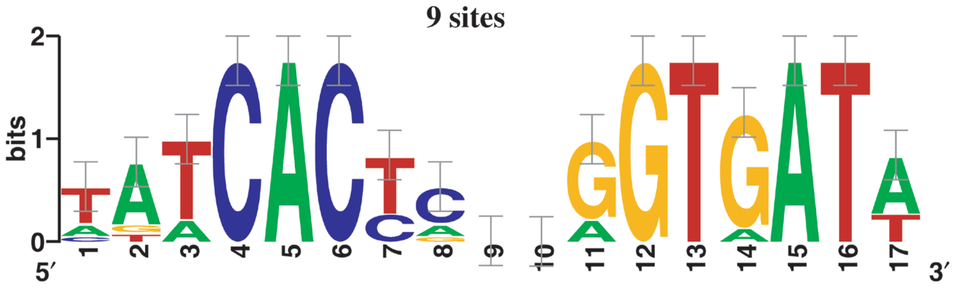
A	7	12	24	23	23	12	8	0	4	4	15	10	25	18
C	12	2	1	0	0	3	4	3	3	0	0	0	0	2
G	8	11	0	0	3	8	8	0	0	0	0	0	2	0
T	0	2	2	4	1	4	7	24	20	23	12	17	0	7

B



A	2	6	1	0	8	0	0	1	2	2	1	0	0	0	7	0	5
C	2	1	1	9	0	8	2	1	1	1	0	0	0	0	0	1	0
G	0	1	0	0	0	0	2	5	2	1	8	9	0	8	2	0	1
T	5	1	7	0	1	1	5	2	4	5	0	0	9	1	0	8	3

C



A	2	7	2	0	9	0	0	2	4	2	2	0	0	1	9	0	6
C	1	0	0	9	0	9	3	6	1	2	0	0	0	0	0	0	0
G	0	1	0	0	0	0	0	1	3	3	7	9	0	8	0	0	0
T	6	1	7	0	0	0	6	0	1	2	0	0	9	0	0	9	3

Figure 4.2 Sequence logos and position-frequency matrices of (**A**) Transcription Factor B-responsive element and TATA-box, based on 27 *P. furiosus* promoter sequences and the Thermococcales-Glycolytic-Motif in promoter sequences, based on 9 glycolytic enzymes in *P. furiosus* (**B**) and *T. kodakaraensis* (**C**). The sequence logos were generated using WebLogo (Crooks *et al.*, 2004).

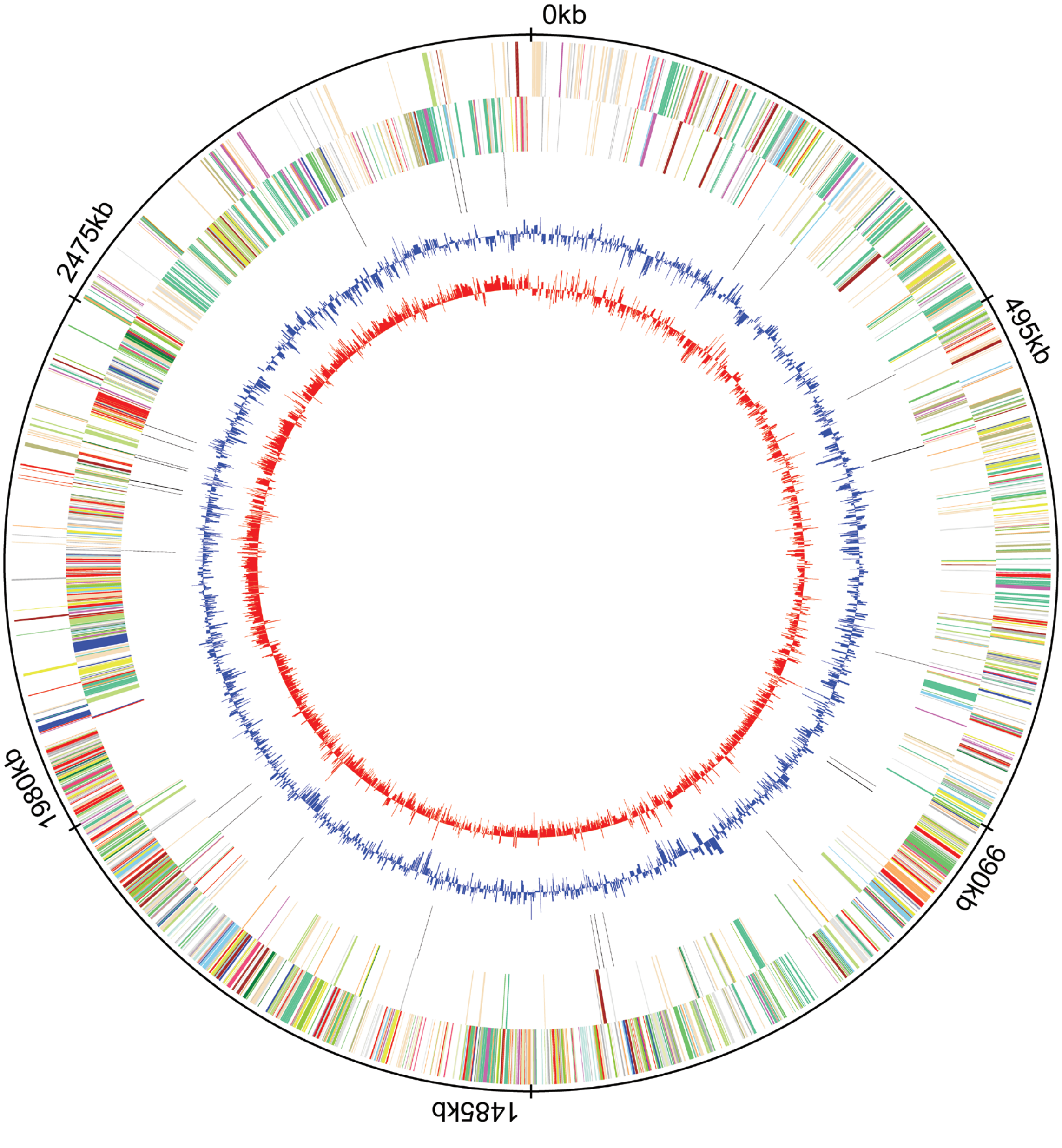


Figure 5.1 Circular representation of the Caldicellulosiruptor saccharolyticus chromosome. From the outer circle to the inner circle (1) genomic position in kilobases (kb) (2) coding sequences on the positive and (3) negative strand, which are colored according to the Clusters of Orthologous Groups of proteins (COG) functional categories, (4) tRNA genes (5) GC% (blue) (5) GC-skew (red). The Microbial Genome Viewer was used to make the circular chromosome wheel (Kerkhoven *et al.*, 2004).

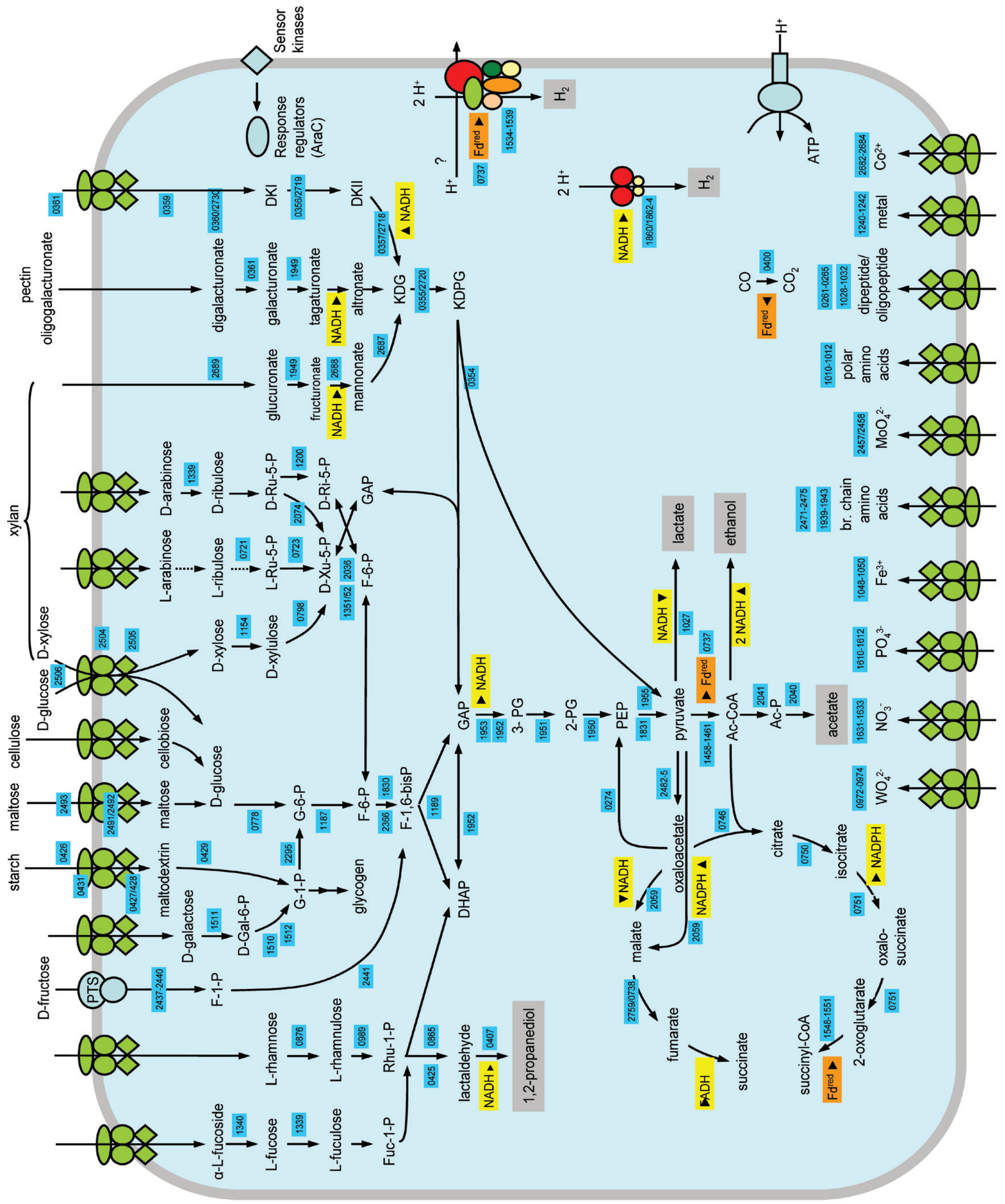


Figure 5.2 An overview of the carbon metabolism and transport systems in *Caldicellulosiruptor saccharolyticus*. The identity of the various ABC-type sugar transporters is not known. Secondary transport systems may be involved as well.

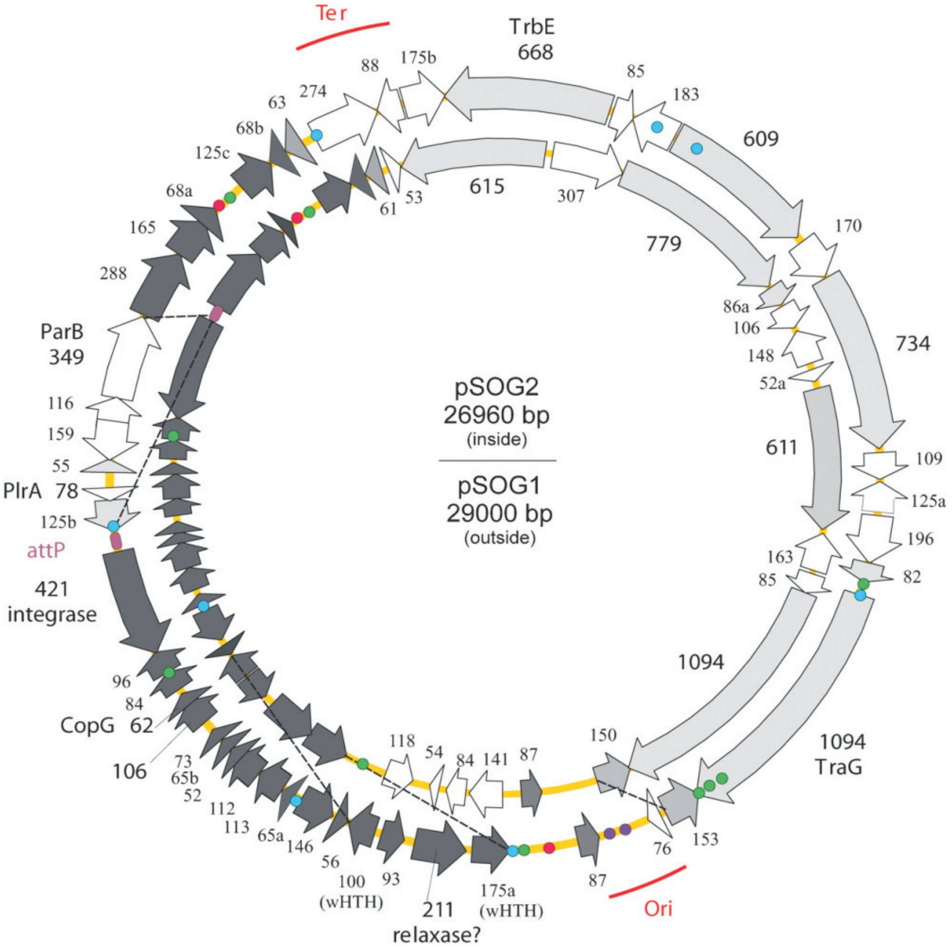


Figure 6.3 Comparison of the pSOG1 and pSOG2 sequences. This diagram shows the circular genomes of pSOG1 on the outside and pSOG2 on the inside. ORFs are shown as arrows. Similar ORFs in the two plasmids are filled in gray; identical ORFs are filled in black; ORFs not conserved between the two plasmids are not filled. ORFs with predicted functions are labelled and ORFs discussed in the text are in bold. Insertions and gene replacements are indicated by dashed lines between the two genomes. ORF names are shown next to the corresponding arrows. The recombination motif TAAACTGGGGAGTTTA is represented by a small disk, colored green when present on the direct DNA strand and light blue when located on the complementary strand. Blue disks indicate the two larger tandem repeats, and a red disk indicates larger inverted repeats. The violet oval represents the putative site of integration attP. The approximate location of the origin (Ori) and terminus (Ter) of replication as predicted by cumulative GC skew and Z-curve analyses are also indicated. This figure was previously published in (Erauso *et al.*, 2006).

About the author

About the author

Harmen Jan George van de Werken was born on 8 September 1971 in ‘s-Hertogenbosch, the Netherlands. In 1989, he completed secondary school, Atheneum B, at the Sint-Janslyceum, ’s-Hertogenbosch and started with the M.Sc. Molecular Sciences at Wageningen University in the same year. During his study he focused on and wrote his M.Sc. theses on the topics applied informatics and microbiology. From 1996-1999, Harmen worked as programmer and technical designer at G&D Software (Capgemini) where he gained experience in managing ICT projects, and was responsible for the implementation of several contract administration at ABN-AMRO. After receiving his M.Sc. degree in Molecular Science in 2002, Harmen was employed by the Fungal Genomics group, Laboratory of Microbiology, Wageningen University, where he developed a web application for administration of chemicals, strains and plasmids. Early 2003, he started his Ph.D. project entitled “Computational Genomics of Prokaryotes”, funded by NWO-BioMolecular Informatics, at the Bacterial Genetics Group, Laboratory of Microbiology, at Wageningen University. The results of this project are presented in this thesis. In January 2008, Harmen was appointed at the Erasmus University Medical Centre, Rotterdam, the Netherlands where he currently has a post-doc position in bioinformatics.

List of publications

van de Werken, H. J. G., Verhees, C. H., Akerboom, J., de Vos, W. M., and van der Oost, J. (2006). Identification of a glycolytic regulon in the archaea *Pyrococcus* and *Thermococcus*. FEMS Microbiol Lett *260*, 69-76.

van de Werken, H. J. G., Verhaart, M. R. A., L., V. A., Willquist, K. U., Lewis, D. L., Nichols, J. D., Goorissen, H. P., Mongodin, E. F., Nelson, K. E., van Niel, E. W. J., Stams, A. J. M., Ward, D. E., de Vos, W. M., van der Oost, J., Kelly, R. M., and Kengen, S. W. M. Hydrogenomics of the extremely thermophilic bacterium *Caldicellulosiruptor saccharolyticus*. submitted for publication.

Snijders, A. P. L., Walther, J., Peter, S., Kinnman, I., de Vos, M. G. J., **van de Werken, H. J. G.**, Brouns, S. J. J., van der Oost, J., and Wright, P. C. (2006). Reconstruction of central carbon metabolism in *Sulfolobus solfataricus* using a two-dimensional gel electrophoresis map, stable isotope labelling and DNA microarray analysis. Proteomics *6*, 1518-1529.

Erauso, G., Stedman, K. M., **van de Werken, H. J. G.**, Zillig, W., and van der Oost, J. (2006). Two novel conjugative plasmids from a single strain of *Sulfolobus*. Microbiology *152*, 1951-1968.

Ortmann, A. C., Brumfield, S. K., Walther, J., McInnerney, K., Brouns, S. J. J., **van de Werken, H. J. G.**, Bothner, B., Douglas, T., van de Oost, J., and Young, M. J. (2008). Transcriptome analysis of infection of the archaeon *Sulfolobus solfataricus* with STIV. J Virol (in press).

Brouns, S. J. J., Walther, J., Snijders, A. P. L., **van de Werken, H. J. G.**, Willemen, H. L. D. M., Worm, P., de Vos, M. G. J., Andersson, A., Lundgren, M., Mazon, H. F. M*., et al.* (2006). Identification of the missing links in prokaryotic pentose oxidation pathways: evidence for enzyme recruitment. J Biol Chem *281*, 27378-27388.

van der Oost, J., Walther, J., Brouns, S. J. J., **van de Werken, H. J. G.**, Snijders, A. P. L., Wright, P. C., Andersson, A., Bernander, R., and de Vos, W. M. (2006). Functional Genomics of the Thermo-Acidophilic Archaeon *Sulfolobus Solfataricus*. In Extremophiles - Methods in Microbiology F. A. Rainey, and A. Oren, eds. (Amsterdam, Elsevier/Academic Press), pp. 201-231.

van de Werken, H. J. G., Brouns, S. J. J., and van der Oost, J. (2008). Pentose Metabolism in Archaea. In Archaea, P. Blum, ed. (Wymondham, Caister Academic Press).

Kanai, T., Akerboom, J., Takedomi, S., **van de Werken, H. J. G.**, Blombach, F., van der Oost, J., Murakami, T., Atomi, H., and Imanaka, T. (2007). A Global Transcriptional Regulator in *Thermococcus kodakaraensis* Controls the Expression Levels of Both Glycolytic and Gluconeogenic Enzyme-encoding Genes. J Biol Chem *282*, 33659-33670.

VLAG graduate school activities

Discipline specific activities

Courses

Prokaryotic Genome Annotation TIGR, Rockville, MD, USA
ERGO course, Nijmegen, 2003

Meetings

Annual Meeting Molecular Genetics, Lunteren, 2003-6 (poster and oral presentations)
Biannual platform meeting NBV, Wageningen, 2003
Symposium ‘Bioinformatics at the Interface’, Utrecht, 2003
European Conference on Prokaryotic Genomes, Göttingen, 2003 (poster)
Wageningen Springschool Bioinformatics, Wageningen, 2004 (poster)
Symposium ‘Images of Life’, Groningen, 2004 (poster)
Gordon research conference Archaea, Oxford UK, 2005 (poster)
Extremophiles, Brest, France, 2006 (poster)
Annual Protein meeting, Oss, 2004

General courses

Scientific writing, Wageningen University, 2005
Organizing and supervising M.Sc. thesis projects, Wageningen University, 2004

Optional courses and activities

VLAG PhD trip, USA, 2006
Bacterial Genetics weekly group Meetings, Wageningen, 2003-2006
Microbiology biweekly group Meetings, Wageningen, 2003-2006
NWO-BMI Computational Genomics project meetings, The Netherlands
Preparation PhD research proposal

Colophon

The research described in this thesis was supported by the Earth and Life Sciences Foundation (ALW) and Chemische Wetenschappen (CW) project 050.50.206, which are subsidized by the Netherlands Organization for Scientific Research (NWO).

Print: Gildeprint drukkerijen BV, Enschede

Cover design: Jos Vogelpoel