Functional and comparative genomics of the Archaea

Thijs J. G. Ettema

Promotoren:

prof. dr. W. M. de Vos
Hoogleraar Microbiologie
Wageningen Universiteit

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

Leden van de
promotiecommissie:

dr. E. V. Koonin
NCBI, NIH, Bethesda, USA

prof. dr. M. A. Huynen
UMC St. Radboud, Nijmegen

prof. dr. S.C de Vries
Wageningen Universiteit

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

Functional and comparative genomics of the Archaea

Thijs J. G. Ettema

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
prof. dr. ir. M.J. Kropff
in het openbaar te verdedigen
op vrijdag 11 november 2005
des namiddags te vier uur in de Aula

Thijs J. G. Ettema – Functional and comparative genomics of the Archaea – 2005
Dutch: 'Functionele en vergelijkende genoomanalyse van de Archaea'

Ph.D Thesis Wageningen University, Wageningen, The Netherlands – With summary in Dutch

ISBN: 90-8504-229-1

Pap en mam,
dit boekje is voor jullie

Cover – Design & Realization: Thijs Ettema, Nico Ettema
Cover picture – Artists Palette, Wai-O-Tapu Thermal Wonderland, New Zealand
Printing – PrintPartners Ipskamp B.V., Nijmegen

DE PROLOOG...

Nu het einde van mijn promotietijd in zicht komt, is de tijd daar om eens terug te blikken op de afgelopen jaren die ik als promotieonderzoeker heb doorgebracht in de Bacteriële Genetica (‘bacgen’) werkgroep van het Laboratorium voor Microbiologie. In het begin, toen nog als student werkend aan enkele afstudeerprojecten, had ik al snel door dat dit voor mij de ideale plek was om een wetenschappelijke loopbaan te beginnen: een gezonde mix van wetenschappelijke diepgang met een fijne sfeer, waar je jezelf op je eigen manier kunt ontplooien. Toen mij vervolgens een AIO-plaats werd aangeboden in de bacgen groep van John van der Oost, was de keuze dan ook snel gemaakt. Nu, 4 mooie, vruchtbare jaren later, heeft dit promotieonderzoek geresulteerd in het proefschrift dat nu voor je ligt... Zoals altijd hebben vele, vele mensen op de een of andere manier bijgedragen aan de totstandkoming van dit proefschrift.

John, jou wil ik allereerst bedanken. Jij hebt mij de mogelijkheid gegeven om dit promotieonderzoek überhaupt uit te voeren. Maar ook gedurende de afgelopen jaren ben je een onophoudelijke bron van inspiratie en motivatie gebleven. Zonder jouw inbreng was dit boekje een stuk dunner geweest. Aan mij de eer om als eerste AIO onder jouw hoogleraarschap te mogen promoveren; hopelijk zullen er nog vele volgen. Willem, als ‘*pater familiaris*’ van Microbiologie ben jij op de achtergrond mijn werk altijd met interesse blijven volgen en zonodig bijsturen. Bedankt hiervoor!

Natuurlijk hebben ook alle collega’s/vrienden van de Bacgen groep, in constant wisselende samenstelling, bijgedragen aan dit proefschrift. Eerst was er de ‘oude garde’ met Corné (mijn oude leermeester), Thijs K/Capri/Kapert/1 (’t was soms wat verwarrend met 2 Thijzen naast elkaar op 1 kamer), Don (when will we have another “one for the road”?), Arjen (gaan we binnenkort nog eens lunchen?), Johan (Giovanni Inguini di Legno) en Ana (je stond goed je mannetje tussen alle mannen!). Leon, als collega, maar later ook als huisgenoot, ik heb veel aan je gehad. Je uitgekookte humor en relativeringsvermogen sleepten me vaak over de dode punten, je zelfgemaakte wijn was van ‘ongekende klasse’! Ook de ‘nieuwe generatie’ bacgenners wil ik graag bedanken. Stan, Jasper (Jaapie), jullie maakten het aangenaam vertoeven in ‘ons hok’. Jullie kweek- alsmede quake capaciteiten worden gemist en de roadtrip naar de Extremophiles in Napels met bijbehorende “@ the crater tour” met Thijs K. behoort zonder twijfel tot de hoogtepunten van mijn promotietijd. Dank gaat ook uit naar Harmen (sterk spul, dat Malt bier), mijn kompaan als het ging om de ‘*in silico* benadering’. Krisztina (my dear companion at Zw & Zn), Odette, Pino, Ronnie, Susan, Hao, Jasper W., Servé (ik zwaai iedere dag als we elkaar voorbijflitsen op de snelweg) en Ans (zonder jou is Bacgen niet Bacgen!): Thanks! Frank (the dude), ons Koffie-ritueel heeft menige frustratie en ochtendhumeur doen verdwijnen; Nico, Mirjam, en andere leden van de Micfys koffie-corner, bedankt voor het aanhoren van mijn ochtendgeraaskal. Nees, Wim, Francis en Jannie, ook jullie hebben op de achtergrond ervoor gezorgd dat ik dagelijks mijn gang kon gaan. Ria, jouw koffie bleek onvervangbaar… Iedereen die ooit heeft meegedaan aan de vele Tour, WK en EK poule's: Bedankt! – Dudes, hopelijk wordt de OLVMTP een langlopende traditie! Stan en Frank, ik ben blij dat jullie als trouwe pinguïns naast me staan vandaag!

Het werk in dit proefschrift had natuurlijk niet tot stand kunnen komen zonder talloze samenwerkingsverbanden. Tina and Hatim, thanks for the fruitful collaboration on deciphering the archaeal ED pathways. Finally we managed to get a solid story! Steve thanks for the experimental advice and for sending those numerous packages to Wageningen. Who knows when it will pay off… Nigel, my TRASH-trip to Newcastle was short but stimulating. Kira and Eugene, I am glad we could combine our efforts to publish our data on the archaeal PEPC. Eugene, I am sincerely honoured that you will be part of my defence committee. Travis and John R., thanks for the collaboration on the RAM-domain paper. Martijn, sinds mijn stage op het EMBL heb ik altijd een beroep kunnen doen op je bioinformatica expertise; jouw input is dan ook terug te vinden in meerdere hoofdstukken van dit proefschrift. Ik hoop dat, nu ik Post-Doc ben in je groep in Nijmegen, we nog een mooie, succesvolle tijd voor de boeg hebben.

Robin (MBF bleek helaas geen 'appeltje-eitje'), Packo (uitvinder van 'loading ladder'), Hinco, Bart, Gera, Noor, Wim, Nico, bedankt voor al die uren bikkelen in het lab. Robin, Bart, hopelijk geeft MBF zijn geheimen ooit prijs…

Omdat hard werken altijd gecompenseerd dient te worden met de nodige ontspanning en vertier, wil ik ook graag mijn familie, vrienden, kennissen en collega's van Micro en Comics bedanken voor de getoonde interesse, de gezellige dagen, avonden, nachten (soms ochtenden…). Bart en Jaap, de broertjesweekenden waren onovertroffen (37!?!); mogen er nog velen volgen! Teamleden van Invicta-H2, bedankt voor het ontspannende beuk en hakwerk, alsmede de 'derde helften'.

Pap en mam, jullie hebben mij de mogelijkheid gegeven om tot *hier* te komen; bedankt voor het vertouwen en de steun, en voor de interesse die jullie altijd hebben getoond in mijn werk (al was het soms moeilijk te verwoorden *wat* ik nu eigenlijk uitspookte al die tijd). Dit boekje is voor jullie.

Kleine Sven, lief kereltje, zo klein, zo pril en zo onschuldig… Door jouw vroege komst was het toch nog 'ff stressen' met dit boekje. Een mooie toekomst ligt voor jou in het verschiet, laat al je dromen uitkomen.

Caroline, mijn allerliefste, tot slot wil ik jou bedanken voor alle kleine èn grote dingen die ertoe doen in het leven, ons leven. Ons kersverse gezinnetje zal altijd het fundament zijn voor liefde en geluk, wanneer dan ook, waar dan ook…

Thijs
November 2005

TABLE OF CONTENTS

Chapter 1 Perspective, aim and outline of this thesis *1*

Chapter 2 Discovering novel biology by *in silico* archaeology *5*

Chapter 3 Modularity in the gain and loss of genes: applications for function prediction *31*

Chapter 4 TRASH: a novel metal-binding domain predicted to be involved in heavy-metal sensing, trafficking and resistance *39*

Chapter 5 A novel type of copper-responsive transcription regulator in archaea *49*

Chapter 6 A novel ligand-binding domain involved in regulation of amino acid metabolism in prokaryotes *71*

Chapter 7 Identification and functional verification of archaeal-type phospho*enol*pyruvate carboxylase, a missing link in archaeal central carbohydrate metabolism *83*

Chapter 8 The semi-phosphorylative Entner-Doudoroff pathway in hyperthermophilic archaea — a re-evaluation *101*

Chapter 9 Summary and concluding remarks *129*

Chapter 10 Nederlandse samenvatting *137*

Appendix I: Colour figures *151*

Curriculum vitae *169*

List of publications *171*

Chapter 1

Perspective, aim and outline of this thesis

In the mid-seventies, Carl Woese and co-workers discovered a class of cellular organisms, referred to as ‘archaebacteria’, that could neither be classified as Bacteria nor Eukaryotes. Later, these organisms were proposed to constitute a separate domain of life: the domain of the Archaea. Early pioneering research on Archaea validated the classification of the Archaea as being a separate prokaryotic domain, and revealed many surprising and unique features. For example, many archaeal species were isolated from hostile habitats at which life had never been expected to be possible, resulting in the nick-name ‘extremophiles’. Decades later, it has become clear that Archaea are not restricted to these hostile environments. On the contrary, they are virtually omnipresent and have been isolated from almost all thinkable habitats, including for instance the rumen of large herbivores and the mammalian gastro-intestinal tract. Current estimates indicate that about 25% of microbial life is of archaeal origin, supporting an important ecological role for archaeal species on earth (e.g. methane production by the metanogens).

However, many aspects with respect to the lifestyle, biology and evolution of the Archaea remained obscure. In the mid-nineties, technologies became available that enabled the sequencing of whole genomes. This period, also referred to as the ‘genomic era’, has revolutionized the developments in molecular biology and shaped the field of the life sciences as we know it today: Instead of looking at genes, scientists were now able to study an organism’s blueprint, its genome.

However, initial studies of completely sequenced genomes revealed many genes for which no function could be assigned. In particular, this appeared to be the case for the Archaea: the first archaeal genome that was completely sequenced, that of *Archaeaoglobus fulgidus*, revealed a mysterious genomescape, for which only for a mere 30 % of all the genes could initially be assigned a function. Clearly, there was a need for rational-based function predictions. Soon, it became apparent that by comparing genomes, it is possible to uncover the valuable clues that reside in them and as such, to increase our insight in function and evolution of a particular biological system: The field of comparative genomics was born, allowing experimentalists to efficiently design an experimental setup for the verification of predicted functions.

The aim of the work that is presented in this thesis was to gain insight in the relatively unknown biology of the Archaea. The approach that was used to achieve this was twofold. First, we have attempted to identify or predict novel functions or functional systems in Archaea, based on the comparison of the available genome sequences (*comparative genomics* approach). Secondly, we have tried to verify some of these predicted functions via an experimental setup (*functional genomics* approach).

Chapter 2 gives an up-to-date review of how comparative analysis of archaeal genomes has been used to improve their annotation, and how it has, often in combination with experimental verification, contributed to the discovery of novel functions.

By comparing genomes, it is also possible to gain insight in their evolution. In **Chapter 3** describes the results of a comparative genome study. By comparing the genome content of three *Pyrococcal* genomes with that of other available archaeal and bacterial genomes, we were able to gain insight in their evolution, in terms of gain and loss of genes. The presented results show how these analyses can be used to refine the prediction of protein function.

Chapter 4 describes the detection TRASH domain. This novel domain was identified when archaeal genomes were analyzed for conserved gene clusters that contained potential transcriptional regulators. One of the identified gene clusters encoded a potential heavy metal resistance regulon, consisting of genes encoding a cation transporter, a potential transcriptional regulator and a metallochaperone-like protein. Close inspection of these proteins revealed a conserved cystein motif, that is potentially involved in metal binding and/or sensing. The novel domain was named TRASH, according to its anticipated function in trafficking, resistance and sensing of heavy metals. **Chapter 5** describes the molecular characterization of the newly identified gene cluster of the thermo-acidophilic crenarchaeon *Sulfolobus solfataricus.* The results of this study describe the first copper-responsive operon in archaea, a new family of archaeal DNA-binding proteins, and support a prominent role of the original TRASH domain in archaeal copper response.

Chapter 6 describes the detection of a novel type of small molecule binding domain, that potentially functions as an allosteric regulatory switch in Archaea and Bacteria. The domain, designated RAM (after regulatory domain of amino acid metabolism), has been found as a fusion with the DNA-binding domain of Lrp-like transcription regulators and with the catalytic domain of some metabolic enzymes. In addition, it is also found as a stand-alone module. The RAM domain is functionally and structurally compared to a related small molecule binding domain, the ACT domain. It is concluded that both domains appear to play analogous roles in controlling key steps in amino acid metabolism at the level of gene expression as well as enzyme activity.

The next two chapters both combine an *in silico* approach with a functional analysis. **Chapter 7** describes the identification of a missing link in archaeal central carbon metabolism, the archaeal phospho*enol*pyruvate carboxylase (atPEPC). By using sensitive sequence comparison methods, we were able to detect a highly conserved, uncharacterized archaeal gene family that was distantly related to the catalytic core of the canonical PEPC, present in Bacteria and Eukarya. Subsequent functional analysis of the representative of this gene family from the hyperthermophilic acidophile *S. solfataricus* confirmed that the encoded protein indeed displayed highly thermostable PEPC activity. In **Chapter 8**, we investigated the modified Entner-Doudoroff (ED) pathway of hyperthermophilic archaea, which, until this research, were generally believed to proceed via a non-phosphorylative variant. A comparative genomics analysis revealed the presence of an ED gene cluster in multiple archaeal genomes. Interestingly, the genes present in the cluster were indicative of a semi- rather than a non-phosphorylative ED variant. In this chapter, we present functional evidence that both modified ED pathways co-exist in hyperthermophiles. This study proves that genome-based comparative approach can provide new information, prompting for a re-evaluation of past knowledge.

Finally, **Chapter 9** contains a concise summary and a reflection of the obtained results. **Chapter 10** shortly describes the results presented in this thesis for the layman in the Dutch language.

Chapter2

Discovering novel biology by *in silico* archaeology

Thijs J. G. Ettema
Willem M. de Vos
John van der Oost

Archaea are prokaryotes that evolved in parallel with bacteria. Since the discovery of the distinct status of the archaea, extensive physiological and biochemical research has been conducted to elucidate the molecular basis of their remarkable lifestyle and their unique biology. Here we discuss how in-depth comparative genomics has been used to improve the annotation of archaeal genomes. Combined with experimental verification, bioinformatics analysis contributes to the ongoing discovery of novel metabolic conversions and control mechanisms, and as such to a better understanding of the intriguing biology of the archaea.

Nature Reviews Microbiology, in press (2005)

ARCHAEA– LIFE, BUT NOT AS WE KNOW IT

It is generally accepted that all forms of cellular life on Earth belong to either the Prokaryotes or the Eukaryotes. Until the mid 1970s, Prokaryotes were considered synonymous to bacteria. In 1977, however, Carl Woese and Gregory Fox have rocked the scientific world by overturning this dogma[8]. By using a new form of genome analysis - a phylogenetic analysis based upon clustering of rRNA sequences - they have suggested the existence of an additional lineage of descend among the Prokaryotes, referred to as the 'archaebacteria'. Their discovery, sometimes referred to as "the Woesean revolution"[9], has triggered a revival of the debate on the origin of life and on

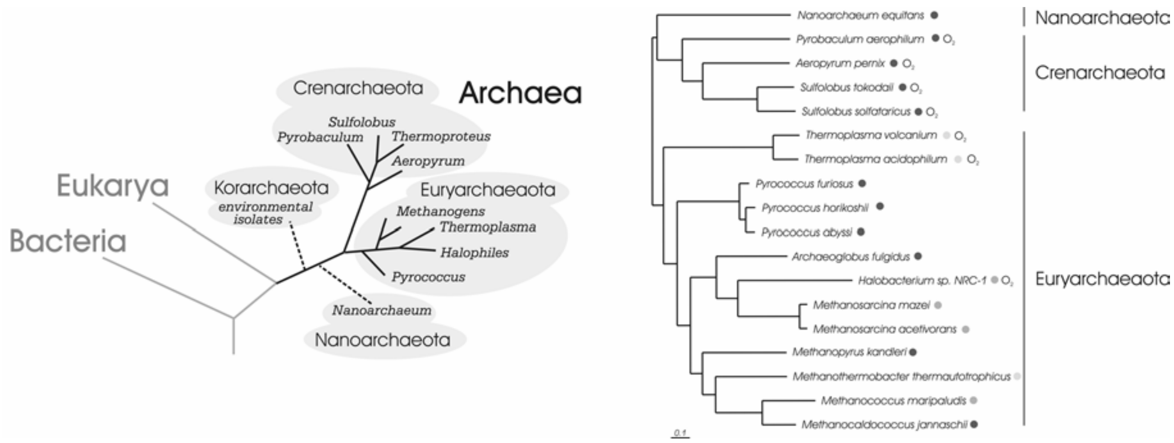


Figure 2.1 Archaeal phylogeny. **(a)** A 16S/18S rDNA-based tree of life (adapted from Woese *et al* [1] and Stetter *et al* [2]) and **(b)** a consensus tree based on a collection of 90 COGs present with a single copy in each archaeal genome (B. Snel, unpublished results). Both trees show that the archaeal domain comprises several phyla. The phylum of the Crenarchaeota is dominated by hyperthermophilic species. Recent evidence, however, indicates that crenarchaea are far from restricted to these hostile environments, as large amount of meso- and psychro-philic crenarchaea have been detected in diverse aqueous[3] and soil samples[4]. The diverse phylum of the Euryarchaeota, contains anaerobic hyperthermophiles (Pyrococcales), thermo-acidophiles (Thermoplasmales), halophiles (Halobacteriales) and the methanogens. The latter group of organisms is also rather diverse and comprises hyperthermophilic (*Methanopyrus, Archaeoglobus*), thermophilic (*Methanothermobacter*), mesophilic (Methanosarcinales) and psychrophilic (Methanogenium) species. The two remaining phyla include the Nanoarchaeota and the Korarchaeota. The former phylum includes one single characterized species, *Nanoarchaeum equitans* [5,6], a hyperthermophilic parasite which relies on its host, the crenarchaeon *Ignicoccus*, as source for lipids, co-factors, amino acids, and nucleotides. Its genome, with only 0.5 Mb the smallest microbial genome sequenced thus far, lacks genes encoding biosynthetic pathways for these essential metabolites[7]. Thus far, members belonging to the phylum of the mysterious Korarchaeota have only been identified by PCR-based amplification of 16S rRNA genes of environmental DNA samples. Red, yellow and blue dots indicate hyperthermophilic, thermophilic and mesophilic species respectively. 'O_2' indicates aerobic species.

the evolution of species. Questions have come up concerning the evolutionary relationship between these 'new organisms' and the domains bacteria and eukarya: do they represent some sort of side branch of the bacteria[10], or rather a new domain of life[1,11]? Consistent with the latter view, Woese has proposed renaming the 'archaebacteria' to the 'archaea', to avoid their mis-classification as 'exotic bacteria' and to stress the three-domain division of life: archaea, bacteria and eukarya (Figure 2.1). Although the unique evolutionary position of the archaea has become widely accepted, the debate on this matter continues[12-14].

ARCHAEA IN THE PRE- AND POSTGENOMIC ERA

Research on archaea in the years after their discovery has validated their classification as a separate, monophyletic prokaryotic domain, and has revealed many surprising and unique features. The archaeal species that Woese and Fox have originally demonstrated to belong to a distinct phylogenetic class, were methane-producing (*Methanobacterium*, *Methanosarcina*), halophilic (*Halobacterium*) and thermo-acidophilic (*Sulfolobus*) micro-organisms[15]. Methanogenic archaea can generally be found in anoxic swamp and lake sediments, and in mammalian gastro-intestinal tracts; in addition, they appear to thrive in geothermal sources and - as endosymbionts - in the cytoplasm of various anaerobic protozoa[16,17]. Initially, most archaeal species have been isolated from habitats where life had never been expected to be possible; archaea from these ecosystems have been named 'extremophiles' (i.e. 'loving extreme conditions'). Apart from the aforementioned microbes that have been isolated from terrestrial sulfur springs (thermo-acidophiles)[2,18] and hypersaline environments (halophiles), archaea have also been found at superheated active volcanic sea floors, often in chimney-like structures called 'black smokers' (hyperthermophiles). Moreover, it has recently become clear that archaea are not restricted to these extreme habitats. They are widespread and abundant in the several diverse niches[19], ranging from the extreme habitats to the human colon and oral cavity[20], and from grassland soils to coalescent microbial communities that live in Antarctic sponges[21]. In addition, significant numbers of archaea have recently been detected in distinct microbial communities, in the course of environmental meta-genome initiatives (e.g in the Sargasso Sea[22], and in an abandoned mine[23]).

Initial biochemical studies also have confirmed the distinct nature of the archaea. For example, archaea differ from bacteria and eukaryotes with respect to their membrane composition, by containing unique ether-lipids rather than ester-lipids[24]. This unique archaeal feature has allowed for a lipid analysis of environmetal samples, revealing that archaea make up as much as 20-30% of the prokaryotic biomass in the oceans[25]. In addition, archaeal cell walls do not contain murein, a polymer that is generally present in bacteria. Further exploration of the molecular biology of the archaea has revealed that they evolved unique systems (lipids, enzymes – see below), but that they certainly also possess typical bacterial and eukaryotic features[26,27]. Being prokaryotes, they are unicellular, lack cell organelles and store their genetic information in a manner that resembles the situation in bacteria: generally a single circular genome with genes that are densely packed in operons. In contrast, archaea contain a eukaryote-like machinery for the processing of genetic information. Most of the protein components that form the basis of the central dogma of molecular biology[28] (the flow of genetic information from DNA to RNA to protein) are orthologous in archaea and eukarya: proteins that drive DNA replication[29-31] [32] (a.o. DNA polymerase, primase and helicases), transcription (basal transcription factors and a complex, multi-subunit RNA polymerase)[33-35] and translation (ribosomes and translation factors)[36,37] are closely related in these domains. In addition, the presence of archaeal counterparts of eukaryal systems involved in chromatin packing and modulation (histones)[38], DNA repair[39], protein turnover (proteasome)[40] and RNA degradation (exosome)[41,42] suggest a common evolutionary origin of these systems. In contrast, they appear to be phylogenetically unrelated to the respective systems that are present in bacteria. Most likely, the eukarya have inherited the genes encoding information processing components from their archaeal ancestor[43,44].

A milestone in the research on Archaea has been the publication of the genome sequence of *Methanocaldococcus jannaschii* in 1996[45]. The first elucidated archaeal genome, the fourth genome ever to be completely sequenced, "revealed the dept of our ignorance of the biology of this remarkable group of organisms"[46]. Initial analysis of the genome has confirmed earlier research suggesting that the majority of the genes in *M. jannaschii* that are linked to energy production, cell division, and metabolism are most similar to counterparts in bacteria, and that most of the genes involved in information processing resemble those in eukaryotes. A more detailed investigation using the homology-based function prediction tools available at the time, has revealed that for only 38% of the genes a reliable functional assignment could be made. Since then, several genomes of phylogenetically and physiologically diverse archaeal species have been sequenced (Table I). The diversity among archaea is also reflected in the

Table I: Sequenced archaeal genomes

	Year of release	OGT [a]	Lifestyle and features	Proteins[b]	Number (%) in COGs[c]	Ref.
Crenarchaeota						
Aeropyrum pernix K1	1999	90	Aerobic chemorganotroph, first sequenced crenarchaeal genome	1,841	1,338 (73)	47
Pyrobaculum aerophilum IM2	2002	100	Facultative nitrate reducing anaerobe	2,605	1,774 (68)	48
Sulfolobus solfataricus P2	2001	87	Aerobic thermo-acidophile, chemorganotroph, sulphur oxidizer, many IS elements	2,976	2,409 (81)	49
Sulfolobus tokodaii str. 7	2001	80	See *S. solfataricus*	2,825	2,147 (76)	50
Euryarchaeota						
Archaeoglobus fulgidus DSM4304	1997	83	Anaerobe, sulfate reducing chemolitho/organo-autotroph	2,420	2,054 (85)	51
Halobacterium sp. NRC-1	2000	50	Aerobic chemorganotroph, obligate halophile, contains two megaplasmids	2,622	2,071 (79)	52
Haloarcula marismortui ATCC 43049	2004	53	See *Halobacterium*, genome encoded by 2 chromosomes and 7 additional plasmids	4,071	2,965 (73)	53
Methanosarcina acetivorans C2A	2002	35	Anaerobic chemolitho (aceto) autotrophic methanogen, nitrogen fixing, forms multi-cellular structures; largest archaeal genome sequenced to date	4,540	3,519 (78)	54
Methanosarcina barkeri str. fusaro	2003	37	See *M. acetivorans*	3,869	3,038 (79)	[b]
Methanosarcina mazei Goe1	2002	37	See *M. acetivorans*	3,371	2,747 (81)	55
Methanocaldococcus jannaschii DSM2661	1996	85	Anaerobic thermophilic chemolithoautotrophic methanogen, two plasmids. First completed archaeal genome sequence	1,785	1,573 (88)	45
Methanocaldococcus maripaludis S2	2004	37	Anaerobic mesophilic chemolithoautotrophic methanogen	1,722	1,530 (89)	56
Methanococcoides burtonii DSM6242	2004	20	Anaerobic methylotrophic marine methanogen, psychrophilic	2,465	2,015 (82)	[b]
Methanopyrus kandleri AV19	2002	110	Anaerobic chemolithoautotrophic methanogen, extreme thermophile, high intracellular salt concentration	1,687	1,353 (80)	57
Methanothermobacter thermoautotrophicus deltaH	1997	65	Anaerobic chemolithoautotrophic methanogen, nitrogen fixing	1,873	1,616 (86)	58
Pyrococcus abyssi GE5	2001	96	Anaerobic heterotrophic thermophile, grows on peptides, contains 1 plasmid	1,895	1,660 (88)	59
Pyrococcus furiosus DSM3638	2001	100	Anaerobic heterotrophic thermophile, grows on peptides *and* sugars	2,125	1,784 (84)	60
Pyrococcus horikoshii OT3	1998	98	See *P. abyssi*	1,955	1,581 (81)	61
Thermococcus kodakaraensis KOD1	2005	95	See *P. furiosus*	2,306	1,850 (80)	62
Ferroplasma acidarmanus fer1	2003	37	Facultative anaerobic acidophile, autotrophic, lacks cell-wall, ferrous-iron-oxidizer	1,872	1,584 (85)	[b]
Picrophilus torridus DSM9790	2004	60	Aerobic thermo-hyperacidophilic heterotroph, grows at pH 0	1,535	1,348 (88)	63
Thermoplasma acidophilum DSM1728	2000	59	Facultative anaerobic chemorganotrophic thermoacidophile, lacks cell-wall	1,481	1,321 (89)	64
Thermoplasma volcanium GSS1	2000	60	See *T. acidophilum*	1,499	1,334 (89)	65
Nanoarchaeaota						
Nanoarchaeum equitans Kin4-M	2003	90	Obligate archaeal parasite, depends on host for metabolites, extremely reduced genome	535	419 (78)	7

[a] Optimal growth temperatures indicated in degrees Celsius (°C) were taken from the Prokaryotic Growth Temperature Database (PGTdb)[66].
[b] Unpublished genome data taken from http://www.ncbi.nlm.nih.gov/genome
[c] Including plasmids

structure and plasticity of their genomes. For example the genome of the methanogen *Methanosarcina acetivorans*[54] (4.5 Mbp; approx. 4,500 genes) apparently has acquired a considerable amount of its genes by horizontal gene transfer from bacteria; in contrast, the genome of the archaeal parasite *Nanoarchaeum equitans* is reduced significantly (0.5 Mbp; approx. 535 genes) and as such appears largely dependent on a tight physical interaction with its specific host[7]. These enormous differences in genome size (and content) are the result of lineage specific gains and losses of genes[67-69] in the course of evolution, and reflect the adaptation of these organisms to a specific environment or niche. The comparison of genomes of closely related archaeal species reveals some interesting examples of genome adaptation. For instance, systematical genome analyses of three *Pyrococcus* species, *P. furiosus*, *P. abyssi* and *P. horikoshii*, reveals significantly more carbohydrate metabolizing genes in the genome of *P. furiosus*, reflecting its adaptation towards a saccharolytic lifestyle[68]. This observation is further supported by the finding by Diruggiero and co-workers, who have provided evidence of a recent horizontal gene transfer event in which *P. furiosus* acquired a 16 kb gene cluster, encoding a maltose/trehalose ABC transporter from a closely related *Thermococcus* strain[70]. In contrast, the genome of *P. horikoshii* has lost many genes encoding enzymes involved in amino acid biosynthesis, reflecting an adaptation towards a proteolytic lifestyle[68,71].

The collection of archaeal genome sequences has created a wealth of data for the archaeal community to study. Initial comparative genome studies on these genomes has revealed a 'conserved archaeal core' of genes[72]. This core, which in later studies has been found to consist of approximately 300 clusters of orthologous groups of proteins (COGs)[46], mainly consists of housekeeping genes involved in information processing; in contrast, it contained only a few genes linked to metabolism. It is suggested that the conserved archaeal core encodes essential functions and therefore has not been subjected to non-orthologous gene displacement ([Glossary term]) during the evolution. In addition to a stable core, archaeal genomes were also found to contain a 'flexible shell' that is sensitive to evolutionary events like lineage specific gain and loss of genes[46]; the flexible shell contains the majority of the metabolic genes, obviously reflecting the wide diversity of archaeal metabolism.

Despite the valuable information that has been derived from archaeal genome sequences; however, many questions relating to functional implications still remain to be answered. As the majority of the genes in most species will never be studied experimentally, our understanding of their biochemistry and physiology will largely rely on the information that is

Box 1 – Homology-based function assignment

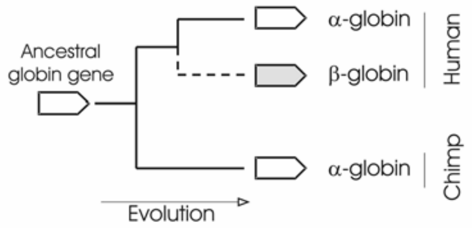
Due to the huge amounts of sequence data being generated from genome sequencing projects during the past decade, only a tiny fraction of all genes and proteins have yet been studied experimentally. Consequently, computational sequence analysis tools are essential for the functional annotation of these genes. In the early days of sequence comparison, the most frequently used techniques were database search algorithms such as BLAST[73] and FASTA[74], designed to identify similar sequences within an estimated statistical significance. Later, optimized tools like PSI-BLAST, gapped BLAST[75] and HMMer[76] have been developed that reduced the computation time needed while simultaneously increasing the sensitivity of a search.

Sometimes, protein sequence comparison is not straightforward, as many proteins can be subdivided into modular moieties, i.e. protein domains. Each domain can exhibit a specific function that may be unrelated to other domains in the same protein (for example DNA binding and ligand sensing in a transcriptional regulator). The domain architecture of proteins can be highly flexible, reflecting the great diversity of modular forms that exist in proteins. This is the result of domain duplication and recombination events, suggesting that protein evolution is a modular process[77-79]. With this in mind, tools and databases have been developed that allow the analysis of domain composition of proteins rather than directly comparing protein sequences. Examples of these domain databases are SMART[80], PFAM[81], CDD[82] and Prosite[83].

The rapidly increasing numbers of gene and protein sequences in databases have also created a need for improved definitions of homologous relationships between proteins, since not all proteins that share homology perform the same molecular function. Extrapolating an assignment of function for a gene/protein from one species to another requires the knowledge of functional counterpart in these species.

The term "orthology" is used to denote this relation between corresponding proteins in an evolutionary sense (Figure 2.2). It refers to homologous proteins of which the independent evolution reflects a speciation event (e.g. α-globin in human and in chimpanzee), in contrast to paralogs, which are the result of a gene duplication event[84,85] (e.g. α-globin and β-globin in human). Orthologous proteins are most likely functional counterparts in various species. The concept of assigning proteins in orthologous clusters has been proven to be very useful in the translation of functional assignments between proteins from different species and has therefore been implemented in databases, e.g. the Clusters of Orthologous Groups of proteins (COG) database (http://www.ncbi.nlm.nih.gov/COG/new/)[86]. However, one should keep in mind that within a defined COG, there might exist proteins that have different physiological functions, e.g. enzymes with a similar catalytic mechanism but with different substrate specificity; this may be reflected by the structure of the dendrogram that is provided for each COG.

Figure 2.2 The concept of orthology. Schematic evolutionary scenario of speciation (solid line) and duplication events (dotted line) of an ancestral globin gene. The α-globin genes in human and chimpanzee (shaded *white*) are orthologs, whereas the genes encoding α-globin (*white*) and β-globin (*gray*) in human have a paralogous relation, since they reflect a duplication event.



derived from experimental data of orthologous systems in model organisms. For a subset of the archaeal genes, the functional prediction is quite straightforward, since the respective proteins contain highly conserved orthologs in bacterial and/or eukaryote model organisms (Box 1). However, for a rather large group of archaeal genes, sequence analysis can only provide an approximate indication of function (e.g. 'probable kinase of the Pfk family'), due to the erosive effect of evolution; elucidation of the function for these groups requires experimental analysis. A major practical limitation is the fact that the development of archaeal genetic systems (allowing gene disruption and/or overexpression) is still in its infancy, making investigation of archaeal gene function not straightforward. Therefore, the established set of conceptually different tools for genomic context-based function prediction (Box 2) are a useful addition to homology-based function prediction that can be used to improve and refine genome annotation[87]. In contrast to homology-based function prediction methods that aim at predicting the specific function of a protein (e.g. enzyme type), context-based function prediction tools tend to predict higher order functions (e.g. enzyme specificity as part of a metabolic process)[88]. Efficient use of these tools can contribute to the *in silico* prediction of novel archaeal functions and facilitate the detection of 'missing links' in archaeal metabolic pathways. The following section will summarize how the different comparative genomics approaches have contributed to our understanding of several aspects of the intriguing biology of the Archaea.

GENOMIC PATTERNS IMPROVE FUNCTIONAL ANNOTATION

Detection of novel and unexpected archaeal systems

The different functional patterns that are mentioned in Box 2 have been shown to be very useful for improving the annotation of generally poorly annotated archaeal genomes. One of the best examples that illustrates the power of comparative genome analysis, is the prediction of the archaeal counterpart of the eukaryote exosome complex by Koonin and co-workers[42]. In eukaryotes, this complex fulfills an essential role in many pathways involved in RNA maturation and degradation. By comparing the gene order in several completely sequenced archaeal genomes a 'super-operon' has been detected that encoded several homologs of the eukaryotic exosome complex, complemented with genes that are believed to encode additional factors that are functionally associated with the complex. They include catalytic subunits

Box 2 – Predicting protein function by comparing genomic context

Genomes harbour valuable information that can be useful for assigning functions to their genes. In prokaryotes, the local genomic context of a gene hints at the possible function of the encoded protein, since genes that are involved in a similar process (e.g. act in the same pathway or biological process, or are part of the same protein complex), tend to be organized in operons[89]. From the early days of genome sequencing, the potential of exploiting genomic context to improve function prediction has been realized[90].

Several conceptually different types of genomic context tools have been implemented for optimizing protein function prediction, each making use of different evolutionary patterns. Initial methods focussed on analysing conserved operon structures to predict protein pairs that interact either physically (part of the same protein complex), or physiologically (part of the same pathway or process)[91-94] (Figure 2.3A, *upper* panel). In addition, a complementary method implemented divergently organized sets of genes, since conserved divergon organization points at transcriptional co-regulation of the corresponding gene products, and thus at their functional relationship[95,96] (Figure 2.3A, *lower* panel). Other methods have been developed that used phylogenetic distribution patterns to infer functional relations between genes. One of these methods uses co-occurrence of genes (Figure 2.3B, *upper* panel) - genes that are functionally related tend to be either present together, or absent together in a given genome. Not surprisingly, this type of context overlaps with the conserved organization of genes (Figure 2.3A), as genes that are functionally related tend to be clustered. Of the subset of genes that co-occur in genomes without clustering, the functional interactions between the corresponding gene products tend to be less dominated by physical interactions[90]. Alternatively, another phylogenetic signal - anti-correlation (Figure 2.3B, lower panel) - can be regarded as a blueprint of non-orthologous gene displacements[97]. This method is based on the complementary occurrence of genes that are not orthologs, but perform the same function. Unlike other *in silico* approaches, this method allows the prediction of a precise functional role of a gene, i.e. the same function as the anti-correlating gene. Finally, gene fusion (Figure 2.3C) is the most direct form of genomic context. The proteins encoded by genes of which homologs are fused tend to have a related function[98], especially if they are orthologs of the

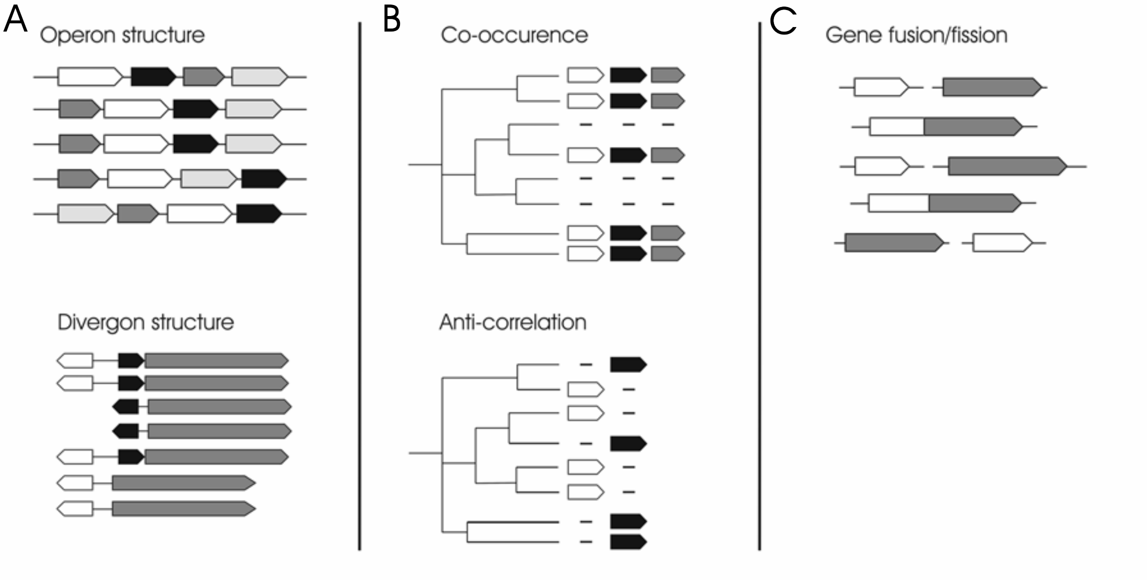


Figure 2.3 (previous page...) Conceptually different types of genomic context that are used for prediction of protein function. **(a)** Conserved local gene context: organization of genes in operon (*upper* panel) and divergon structures (*lower* panel) indicate a functional interaction between the encoded proteins, in for instance biochemical pathways or protein complexes. **(b)** Phylogenetic patterns across genomes: co-occurrence (*upper* panel) and anti-correlation patterns of genes (*lower* panel). The latter type of genomic context reflects non-orthologous gene displacements. **(c)** Gene fusion and fission events are the most direct type of genomic context and point at physical interactions of the encoded proteins.

of the archaeal proteasome, two ribosomal proteins and a DNA-directed RNA polymerase subunit. These observations suggest that in Archaea, a tight functional coupling exists between transcription, translation, RNA processing and degradation, and protein degradation[42]. Recently, the existence of the predicted archaeal exosome-like complex has indeed been demonstrated experimentally in *S. solfataricus*[41].

Some unique features of the archaeal translation machinery have recently been uncovered as well. Analysis of the earlier sequenced archaeal genomes has revealed that some of the genes encoding aminoacyl-tRNA synthetases are missing, including lysyl-tRNA synthetase (LysRS) and cysteinyl-tRNA synthetase (CysRS). Via reverse-genetics approaches, a new class of LysRS has been discovered[101]. Moreover, in stead of a classical CysRS, two novel archaeal enzymes have recently been described: a tRNA synthetase catalyzing the coupling of phospho-serine (Sep) to tRNA^{Cys}, and a synthase responsible for the subsequent conversion of Sep-tRNA^{Cys} to Cys-tRNA^{Cys}[102]. Comparative genomics analysis has revealed that the latter system is present in all organisms that lack CysRS[102]. These examples indicate that the evolution of aminoacyl-tRNA synthesis in many instances is not orthologous as was previously assumed[103]. Yet another interesting case showing the versatility of the archaeal translation machinery has recently been discovered in the genome of the parasitic hyperthermophile *Nanoarchaeum equitans*. Initial analysis of this genome has failed to identify genes encoding the glutamate, histidine and initiator methionine tRNA species. Using an algorithm designed to recognize tRNA signature sequences, Söll and co-workers have successfully identified the genes that encode either the 5′ or the 3′ halves of the missing tRNA genes[102,104]. Interestingly, the site of truncation is just downstream the anticodon, a conserved location for introns in many full-length tRNA genes in other organisms. Experimental analysis has indeed demonstrated the presence as well as the aminoacylation of the predicted full-sized tRNA^{His} and tRNA^{Glu}. The tRNA halves apparently anneal via a 12-14 nucleotide GC-rich RNA duplex, and are subsequently processed by a hitherto unknown mechanism.

Two species of $tRNA^{Glu}$ are potentially encoded by 3 half genes: two 5´fragments with different anticodons CUC and UUC, and a single 3´ fragment; in this case the truncation could be considered as an efficiency solution for this minimal genome[102]. Apart from these split tRNA genes, *N. equitans* has been reported to contain a high number of split protein-encoding genes[7]. Although it remains unclear if why *N. equitans* would benefit from gene fission, it has previously been reported by Snel and co-workers that split genes are slightly overrepresented in the genomes of hyperthermophiles[100].

Analysis of conserved gene context in prokaryotic genomes has also revealed a previously undetected DNA repair system, consisting of more than 20 genes. The repair system has been found to be present in most archaeal genomes and in some thermophilic bacteria[105]. The gene composition and gene order in the conserved context appears to vary greatly between species, but all versions contain a stable, conserved core that consisted of five genes. The genes that make up the core of the DNA repair system include a predicted DNA helicase, an exonuclease of the RecB family, and three uncharacterized genes that potentially encode a novel type of nuclease[105]. Recently, a protein with nuclease-ATPase activity, designated Nar71, has been isolated from *M. thermautotrophicus* cell-extracts as part of the predicted archaeal DNA repair system; its nuclease activity appears to be specific for single-stranded DNA regions. Hence, this finding supports the presence of novel DNA repair functions within the proposed DNA repair gene neighbourhood[106].

In addition, conserved gene context has been used to predict target genes for transcriptional regulators. An analysis of conserved divergently-organized gene pairs in bacterial genomes revealed that in most instances, one of the genes encodes a transcriptional regulator and the other encodes a non-regulatory protein[95]. Subsequently, an in-depth analysis in which gene expression data were integrated with the occurrence of divergent gene pairs in *E. coli*, has indicated that these regulators tend to co-regulate the expression of the divergently organized target gene/operon as well as their own expression. A similar approach to analyse archaeal genomes has revealed the presence of an archaeal heavy metal resistance cluster (Figure 2.4A)[107,108]. The cluster has been shown to consist of a novel archaea-specific transcriptional regulator, a P-type cation transporting ATPase and a predicted metallochaperone, all of which contain a novel type of metal binding domain - the TRASH domain[108]. It has been found that the transcriptional regulator is divergently positioned with respect to the genes encoding the metallochaperone and cation transporter (Figure 2.4A). In a subsequent study of the novel metal resistance cluster in *S. solfataricus*, it has been confirmed that the transcriptional regulator is indeed involved in the transcription regulation of these genes[107].

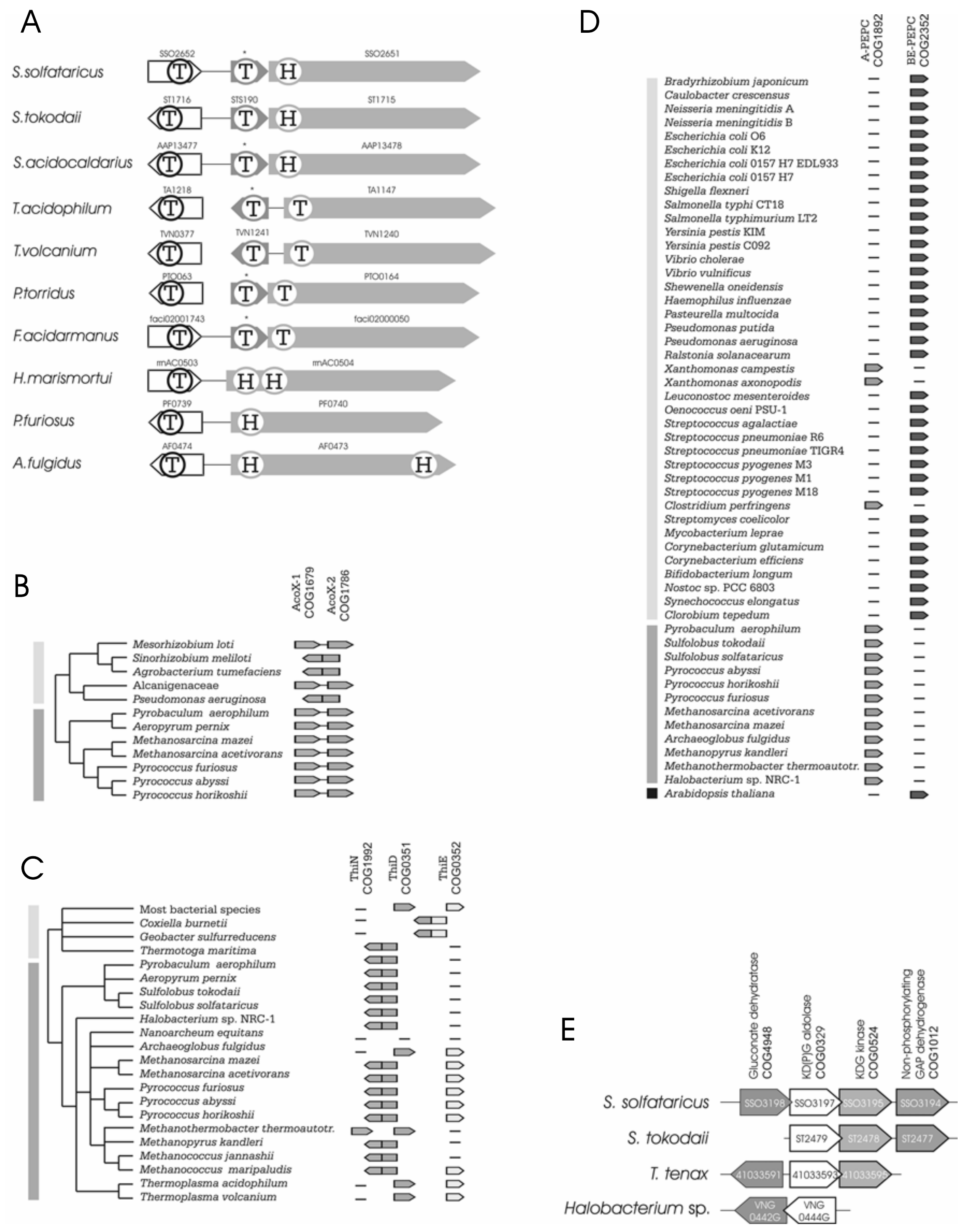


Figure 2.4 (previous page…) Detection of novel functions in archaea using different types of genomic context. **(a)** Conserved gene organization of archaeal *cop* clusters. Genes encoding CopT, CopM and CopA are indicated in *white*, *dark gray* and *light gray* respectively. Drawn lines indicate a physical linkage between genes. Gene numbers are indicated above the genes. Genes that have been overlooked in genome annotation and that are absent from protein databases are indicated with an asterisk. Genes are approximately drawn to scale. Abbreviations of metal binding domain: *T*: TRASH domain; *H*: HMA domain. **(b)** Prediction of the archaeal HMP di-kinase (ThiN, COG1992) based on complementary phylogenetic patterns between COG1992 and the canonical HMP di-kinase (ThiE, COG0352) and multiple gene fusion events with ThiD (COG0351). **(c)** The archaeal aconitase (AcoX) is predicted to comprise two different subunits (COG1679 and COG1786), which most likely form a single complex based on conserved operon organization and multiple gene fusion events in bacterial genomes. **(d)** Complementary pattern of genes encoding phospho*enol*pyruvate carboxylase (PEPC), the archaeal type (COG1892, *light gray*) and the bacterial/eukaryote type (COG2352, dark gray) across different genomes. Genomes of organisms lacking a PEPC encoding gene are not shown. **(e)** Analysis of conserved Entner-Doudoroff gene clusters en coded by archaeal genomes. The presence of a 2-keto-3-deoxy gluconate kinase (KDG kinase, COG0524) suggests that the degradation of glucose in these archaea proceeds (at least partly) via a semi-phosphorylative ED pathway rather than exclusively via the non-phosphorylative version. Orthologous groups of the key genes of the semi-phosphorylative ED are indicated in the same colour (adapted from Achmed *et al*[109]). Genes are indicated by their systematic gene name. Orthologous genes are indicated in the same colour. *Light grey*, *dark grey* and *black* bars indicate bacterial, archaeal and eukaryote species respectively. Physically linked genes are connected with a line.

Filling in gaps in archaeal pathways

An example that illustrates the efficacy of analysis of conserved gene context to fill in gaps that exist in several archaeal metabolic pathways is the identification of the archaeal shikimate kinase, an AAA-type kinase that is related to adenylate kinase, cytidylate kinase and gluconate kinase. Shikimate kinase is an essential enzyme in the biosynthesis of aromatic amino acids and many other aromatic compounds. Initial analysis of archaeal genomes has failed to identify a candidate gene encoding shikimate kinase. However, by analyzing conserved clusters of chorismate biosynthetic genes encoded by archaeal genomes, a gene encoding a predicted kinase of the GHMP superfamily has been identified that could account for the missing archaeal shikimate kinase. A subsequent experimental investigation has confirmed this prediction[110].

The successful allocation of the archaeal aconitase, a missing link in archaeal central metabolism, is a good example of a study in which different types of genomic context have been integrated. Aconitase - an essential enzyme of the tricarboxylic acid cycle[54] - is only represented in few archaeal genomes. Using iterative sequence database searches, two previously uncharacterized protein families (COG1679 and COG1786) were predicted to comprise the missing aconitase domains in archaea and some proteobacteria.

The prediction of the archaeal aconitase has further been supported by genome context: the respective genes are often found in one predicted operon and are fused in several species (Figure 2.4B). The latter finding is a strong indication for a functional and physical interaction between the respective proteins[111]. In addition, two missing links in archaeal glycolysis, the archaeal fructose-1,6-bisphosphate aldolase and phosphoglycerate mutase, have first been predicted by sensitive sequence analysis (PSI-Blast)[112] and have later been verified by laboratory experiments[113,114].

A similar example where integration of different types of genomic context has contributed to a reliable functional prediction is the identification of the archaeal-type phospho*enol*pyruvate carboxylase, PEPC[115]. Despite the fact that archaeal PEPC activity has been reported, the corresponding gene has never been identified. Iterative sequence analysis has revealed distant homology between the bacterial/eukaryote-type PEPC (COG2352) and an uncharacterized protein family (COG1892) that is encoded by most archaeal genomes. The perfectly anti-correlating phylogenetic pattern (Box 2, Figure 2.3B) of the PEPC-containing COGs (Figure 2.4C) has greatly enhanced the reliability of the prediction, and suggests that the presence of the archaeal PEPC in some bacterial genomes is the result of a non-orthologous gene displacement. Sequence comparison with the bacterial/eukaryote-type enzyme has revealed that the archaeal PEPC does contain the catalytic core, but that it lacks the domain responsible for allosteric regulation. Subsequent cloning and functional expression of the COG1892 representative of *S. solfataricus* has indeed confirmed the prediction of an archaeal-type PEPC[115].

The archaeal thiamin biosynthesis pathway is still not well understood. While some enzymatic steps in this pathway are encoded by archaeal genomes, others enzymes appear to be absent. In many instances, missing enzymes are displaced by analogs, functionally equivalent proteins that have evolved independently and lack sequence and structural similarity. Analysis of anti-correlating gene patterns (Box 2, Figure 2.3B), has identified an archaeal candidate for the missing 2-methyl-4-amino-5-hydroxymethylpyrimidine kinase (HMP di-kinase), which is encoded by most archaeal genomes and by the genome of *Thermotoga maritima*[116,117]. The archaeal gene family, designated *thi*N (COG1992), displays an almost perfect anti-correlation with the canonical *thi*E (COG0352), and its involvement in thiamine biosynthesis has further been supported by the observed fusion of orthologs of *thi*N with the gene encoding the bi-functional enzyme ThiD (which catalyzes both phosphorylation steps from 2-methyl-4-amino-5-hydroxymethylpyrimidine, COG0351) in many species (Figure 2.4D). It has therefore been postulated that the 'ThiDN' fusion protein might catalyze the three subsequent steps of the archaeal thiamine biosynthesis pathway[117]. Indeed, the archaeal *thi*N has been demonstrated to functionally complement an *E. coli thi*E deletion strain.

Adjusting and refining assigned functions based on analysis of genome context data

As stated above, context-based tools often only produce higher order function predictions, in contrast to homology-based function prediction, that tends to predict the molecular function of gene products. Sometimes, combining context-based and comparative genomics data allows for refinement of prediction of protein function and functional interactions.

An example nicely demonstrating this is the functional reassignment of RNAse L inhibitor protein (RLI1, COG1245) by Huynen and Gabaldon[118]. This protein is encoded by all archaeal and eukaryal genomes sequenced thus far. Human RLI1 has been implicated in an interferon induced RNA degradation pathway (2'-5' oligoadenylate pathway), where RLI1 inhibits RNase L activity by interacting with this protein, and as such inhibiting the degradation of certain viral RNA species[119]. In addition, RLI1 appears to be involved in capsid assembly of HIV-1[120]. However, the function of at least the archaeal RLI1 orthologs is most likely different since archaeal genomes do not encode RNase L homologs; at presence there is no evidence on a link of the archaeal RLI1 with viral development. To investigate a possible consensus role of RLI1 in archaea and eukarya, orthologous groups of proteins with a similar phylogenetic profile (present in all archaeal and eukaryal genomes, absent from all bacterial genomes) have been examined. This analysis has retrieved 55 orthologous groups, of which the vast majority is either involved in translation and ribosome biogenesis, as well as in other core processes like transcription, and DNA replication, recombination and repair. The observed correlation clearly implicates a role for RLI1 in DNA replication/transcription and/or RNA processing. The anticipated function of RLI1 could even be further specified by integrating additional types of genomics data. Analysis of conserved co-expression data revealed that in eukaryotes, RLI1 is co-expressed with genes involved in the ribosome and ribosome biogenesis, specifically in processing of ribosomal RNA. Furthermore, analysis of protein-protein interaction data has revealed that in yeast, RLI1 specifically interacts with HCR1, a protein that is involved in processing of ribosomal RNA. Also, down-regulation of RLI1 expression appears to correlate with reduced levels of protein biosynthesis[121], which again is consistent with a role in ribosome biogenesis. Finally, analysis of the domain composition of RLI1 hinted at an interaction between the N-terminal four-cysteine (potentially [FeS]) domains of RLI1 with the backbone of rRNA. Indeed, recent publications have confirmed the involvement of RLI1 in ribosome biogenesis[122,123] and have indicated that *rli*1 mutants are defective in pre-rRNA processing. Furthermore, RLI1 has been found to be associated with premature and mature 40S ribosomal subunits in yeast. It is anticipated that RLI1 fulfills similar roles in archaea.

Sometimes, analysis of conserved gene context does result in surprising findings, prompting a re-evaluation of assumptions that have been taken for granted for many years. In our ongoing attempts to study archaeal central carbon metabolic genes and pathways, we have detected a partly conserved Entner-Doudoroff (ED) gene cluster (Figure 2.4E) in several hyperthermophilic archaea. Following an initial exploration of archaeal central metabolism, a non-phosphorylative ED pathway has been proposed to operate in *Sulfolobus*[124]; this pathway typically lacks the phosphorylation of C6-intermediates of this glycolytic pathway. Therefore, the presence of a gene encoding a carbohydrate kinase of the PfkB family in this conserved ED gene cluster was an unexpected finding. In follow-up experiments, it has been shown that the putative carbohydrate kinase in fact displays 2-keto-3-deoxy gluconate kinase activity, suggesting that the degradation of glucose proceeds (at least partly) via a semi-phosphorylative ED pathway in these archaeal thermophiles[109].

FUTURE CHALLENGES AND PERSPECTIVES

The examples described above demonstrate the power of applying a set of conceptually different comparative genomics methods to fill in the ‘black holes’ in our understanding of archaeal metabolism, and as such to increase our insight in archaeal biology. However, these methods still fail to predict molecular functions for a significant subset of the archaeal genes, e.g. those genes annotated as ‘conserved hypothetical protein’. It is clear that the function of these genes can only be elucidated by laboratory experimentation. For this purpose, the development and optimization of efficient experimental tools for some selected archaeal model systems will be an essential step.

Recent advances in function prediction have focused on the analysis of large scale functional genomics datasets including gene expression data (e.g. SAGE[125] and micro-array technology), proteomics data (e.g. 2D electrophoresis, ICAT or stable isotope labeling, mass spectrometry), as well as protein-protein interaction data (e.g. yeast 2-hybrid analysis[126], TAPtagging[38] and mass spectrometry analysis of protein complexes). Initial attempts to integrate these holistic, high-throughput systems biology approaches have already led to novel insights in model organisms like *E. coli*, *B. subtilis*, yeast and also human (e.g. see refs 127, 128 and 129).

It is obvious that a further development of the aforementioned integrative systems biology approaches will be crucial in order to advance our understanding of the intriguing biology of archaea. The recent establishment of several archaeal functional genomics initiatives (e.g. see ref 130) is anticipated to lead to the discovery of many more 'archaeological' treasures.

ACKNOWLEDGEMENT

We thank B. Snel (CMBI, NCMLS, Radboud University Nijmegen Medical Centre, The Netherlands) for construction of Figure 1b.

REFERENCES

1 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**, 4576-9

2 Stetter, K.O. (1999) Extremophiles and their adaptation to hot environments, *FEBS Lett* **452**, 22-5

3 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

4 Ochsenreiter, T., Selezi, D., Quaiser, A., Bonch-Osmolovskaya, L. and Schleper, C. (2003) Diversity and abundance of Crenarchaeota in terrestrial habitats studied by 16S RNA surveys and real time PCR, *Environ Microbiol* **5**, 787-97

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

6 Huber, H., Hohn, M.J., Stetter, K.O. and Rachel, R. (2003) The phylum Nanoarchaeota: present knowledge and future perspectives of a unique form of life, *Res Microbiol* **154**, 165-71

7 Waters, E., Hohn, M.J., Ahel, I., Graham, D.E., Adams, M.D., Barnstead, M., Beeson, K.Y., Bibbs, L., Bolanos, R., Keller, M., Kretz, K., Lin, X., Mathur, E., Ni, J., Podar, M., Richardson, T., Sutton, G.G., Simon, M., Soll, D., Stetter, K.O., Short, J.M. and Noordewier, M. (2003) The genome of *Nanoarchaeum equitans*: insights into early archaeal evolution and derived parasitism, *Proc Natl Acad Sci U S A* **100**, 12984-8

8 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**, 5088-90

9 Pennisi, E. (2004) Evolutionary biology. The birth of the nucleus, *Science* **305**, 766-8

10 Woese, C.R. and Gupta, R. (1981) Are archaebacteria merely derived 'prokaryotes'? *Nature* **289**, 95-6

11 Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magrum, L.J., Zablen, L.B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B.J., Stahl, D.A., Luehrsen, K.R., Chen, K.N. and Woese, C.R. (1980) The phylogeny of prokaryotes, *Science* **209**, 457-63

12 Woese, C.R. (1998) Default taxonomy: Ernst Mayr's view of the microbial world, *Proc Natl Acad Sci U S A* **95**, 11043-6

13 Mayr, E. (1998) Two empires or three? *Proc Natl Acad Sci U S A* **95**, 9720-3

14 Rivera, M.C. and Lake, J.A. (1992) Evidence that eukaryotes and eocyte prokaryotes are immediate relatives, *Science* **257**, 74-6

15 Woese, C.R., Magrum, L.J. and Fox, G.E. (1978) Archaebacteria, *J Mol Evol* **11**, 245-51

16 Embley, T.M., Finlay, B.J., Thomas, R.H. and Dyal, P.L. (1992) The use of rRNA sequences and fluorescent probes to investigate the phylogenetic positions of the anaerobic ciliate Metopus palaeformis and its archaeobacterial endosymbiont, *J Gen Microbiol* **138**, 1479-87

17 Vogels, G.D. and Stumm, C. (1980) Interactions between methanogenic bacteria and hydrogenic ciliates in the rumen, *Antonie Van Leeuwenhoek* **46**, 108

18 Segerer, A.H., Burggraf, S., Fiala, G., Huber, G., Huber, R., Pley, U. and Stetter, K.O. (1993) Life in hot springs and hydrothermal vents, *Orig Life Evol Biosph* **23**, 77-90

19 DeLong, E.F. and Pace, N.R. (2001) Environmental diversity of bacteria and archaea, *Syst Biol* **50**, 470-8

20 Lepp, P.W., Brinig, M.M., Ouverney, C.C., Palm, K., Armitage, G.C. and Relman, D.A. (2004) Methanogenic Archaea and human periodontal disease, *Proc Natl Acad Sci U S A* **101**, 6176-81

21 Webster, N.S., Negri, A.P., Munro, M.M. and Battershill, C.N. (2004) Diverse microbial communities inhabit Antarctic sponges, *Environ Microbiol* **6**, 288-300

22 Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H. and Smith, H.O. (2004) Environmental genome shotgun sequencing of the Sargasso Sea, *Science* **304**, 66-74

23 Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S. and Banfield, J.F. (2004) Community structure and metabolism through reconstruction of microbial genomes from the environment, *Nature* **428**, 37-43

24 Hanford, M.J. and Peeples, T.L. (2002) Archaeal tetraether lipids: unique structures and applications, *Appl Biochem Biotechnol* **97**, 45-62

25 Schouten, S., Hopmans, E.C., Pancost, R.D. and Damste, J.S. (2000) Widespread occurrence of structurally diverse tetraether membrane lipids: evidence for the ubiquitous presence of low-temperature relatives of hyperthermophiles, *Proc Natl Acad Sci U S A* **97**, 14421-6

26 Doolittle, W.F. and Logsdon, J.M., Jr. (1998) Archaeal genomics: do archaea have a mixed heritage? *Curr Biol* **8**, R209-11

27 Koonin, E.V., Mushegian, A.R., Galperin, M.Y. and Walker, D.R. (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the archaea, *Mol Microbiol* **25**, 619-37

28 Crick, F. (1970) Central dogma of molecular biology, *Nature* **227**, 561-3

29 Dionne, I., Robinson, N.P., McGeoch, A.T., Marsh, V.L., Reddish, A. and Bell, S.D. (2003) DNA replication in the hyperthermophilic archaeon *Sulfolobus solfataricus*, *Biochem Soc Trans* **31**, 674-6

30 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**, 7046-51

31 Robinson, N.P., Dionne, I., Lundgren, M., Marsh, V.L., Bernander, R. and Bell, S.D. (2004) Identification of two origins of replication in the single chromosome of the archaeon *Sulfolobus solfataricus*, *Cell* **116**, 25-38

32 Myllykallio, H., Lopez, P., Lopez-Garcia, P., Heilig, R., Saurin, W., Zivanovic, Y., Philippe, H. and Forterre, P. (2000) Bacterial mode of replication with eukaryotic-like machinery in a hyperthermophilic archaeon, *Science* **288**, 2212-5

33 Bell, S.D. and Jackson, S.P. (1998) Transcription in Archaea, *Cold Spring Harb Symp Quant Biol* **63**, 41-51

34 Huet, J., Schnabel, R., Sentenac, A. and Zillig, W. (1983) Archaebacteria and eukaryotes possess DNA-dependent RNA polymerases of a common type, *Embo J* **2**, 1291-4

35 Thomm, M. (1996) Archaeal transcription factors and their role in transcription initiation, *FEMS Microbiol Rev* **18**, 159-71

36 Kyrpides, N.C. and Woese, C.R. (1998) Archaeal translation initiation revisited: the initiation factor 2 and eukaryotic initiation factor 2B alpha-beta-delta subunit families, *Proc Natl Acad Sci U S A* **95**, 3726-30

37 Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution, *Science* **289**, 905-20

38 Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G. and Superti-Furga, G. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes, *Nature* **415**, 141-7

39 White, M.F. (2003) Archaeal DNA repair: paradigms and puzzles, *Biochem Soc Trans* **31**, 690-3

40 Baumeister, W. and Lupas, A. (1997) The proteasome, *Curr Opin Struct Biol* **7**, 273-8

41 Evguenieva-Hackenburg, E., Walter, P., Hochleitner, E., Lottspeich, F. and Klug, G. (2003) An exosome-like complex in *Sulfolobus solfataricus*, *EMBO Rep* **4**, 889-93

42 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**, 240-52

43 Rivera, M.C. and Lake, J.A. (2004) The ring of life provides evidence for a genome fusion origin of eukaryotes, *Nature* **431**, 152-5

44 Lake, J.A. (1989) Origin of the eukaryotic nucleus: eukaryotes and eocytes are genotypically related, *Can J Microbiol* **35**, 109-18

45 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., Kerlavage, A.R., Dougherty, B.A., Tomb, J.F., Adams, M.D., Reich, C.I., Overbeek, R., Kirkness, E.F., Weinstock, K.G., Merrick, J.M., Glodek, A., Scott, J.L., Geoghagen, N.S. and Venter, J.C. (1996) Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*, *Science* **273**, 1058-73

46 Makarova, K.S. and Koonin, E.V. (2003) Comparative genomics of Archaea: how much have we learned in six years, and what's next? *Genome Biol* **4**, 115

47 Kawarabayasi, Y., Hino, Y., Horikawa, H., Yamazaki, S., Haikawa, Y., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Nakazawa, H., Takamiya, M., Masuda, S., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Kikuchi, H. and et al. (1999) Complete genome sequence of an aerobic hyper-thermophilic crenarchaeon, *Aeropyrum pernix* K1, *DNA Res* **6**, 83-101, 145-52

48 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**, 984-9

49 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., Erauso, G., Fletcher, C., Gordon, P.M., Heikamp-de Jong, I., Jeffries, A.C., Kozera, C.J., Medina, N., Peng, X., Thi-Ngoc, H.P., Redder, P., Schenk, M.E., Theriault, C., Tolstrup, N., Charlebois, R.L., Doolittle, W.F., Duguet, M., Gaasterland, T., Garrett, R.A., Ragan, M.A., Sensen, C.W. and Van der Oost, J. (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2, *Proc Natl Acad Sci U S A* **98**, 7835-40

50 Kawarabayasi, Y., Hino, Y., Horikawa, H., Jin-no, K., Takahashi, M., Sekine, M., Baba, S., Ankai, A., Kosugi, H., Hosoyama, A., Fukui, S., Nagai, Y., Nishijima, K., Otsuka, R., Nakazawa, H., Takamiya, M., Kato, Y., Yoshizawa, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K., Masuda, S., Yanagii, M., Nishimura, M., Yamagishi, A., Oshima, T. and Kikuchi, H. (2001) Complete genome sequence of an aerobic thermoacidophilic crenarchaeon, *Sulfolobus tokodaii* strain7, *DNA Res* **8**, 123-40

51 Klenk, H.P., Clayton, R.A., Tomb, J.F., White, O., Nelson, K.E., Ketchum, K.A., Dodson, R.J., Gwinn, M., Hickey, E.K., Peterson, J.D., Richardson, D.L., Kerlavage, A.R., Graham, D.E., Kyrpides, N.C., Fleischmann, R.D., Quackenbush, J., Lee, N.H., Sutton, G.G., Gill, S., Kirkness, E.F., Dougherty, B.A., McKenney, K., Adams, M.D., Loftus, B., Venter, J.C. and et al. (1997) The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*, *Nature* **390**, 364-70

52 Ng, W.V., Kennedy, S.P., Mahairas, G.G., Berquist, B., Pan, M., Shukla, H.D., Lasky, S.R., Baliga, N.S., Thorsson, V., Sbrogna, J., Swartzell, S., Weir, D., Hall, J., Dahl, T.A., Welti, R., Goo, Y.A., Leithauser, B., Keller, K., Cruz, R., Danson, M.J., Hough, D.W., Maddocks, D.G., Jablonski, P.E., Krebs, M.P., Angevine, C.M., Dale, H., Isenbarger, T.A., Peck, R.F., Pohlschroder, M., Spudich, J.L., Jung, K.W., Alam, M., Freitas, T., Hou, S., Daniels, C.J., Dennis, P.P., Omer, A.D., Ebhardt, H., Lowe, T.M., Liang, P., Riley, M., Hood, L. and DasSarma, S. (2000) Genome sequence of *Halobacterium* species NRC-1, *Proc Natl Acad Sci U S A* **97**, 12176-81

53 Baliga, N.S., Bonneau, R., Facciotti, M.T., Pan, M., Glusman, G., Deutsch, E.W., Shannon, P., Chiu, Y., Weng, R.S., Gan, R.R., Hung, P., Date, S.V., Marcotte, E., Hood, L. and Ng, W.V. (2004) Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea, *Genome Res* **14**, 2221-34

54 Galagan, J.E., Nusbaum, C., Roy, A., Endrizzi, M.G., Macdonald, P., FitzHugh, W., Calvo, S., Engels, R., Smirnov, S., Atnoor, D., Brown, A., Allen, N., Naylor, J., Stange-Thomann, N., DeArellano, K., Johnson, R., Linton, L., McEwan, P., McKernan, K., Talamas, J., Tirrell, A., Ye, W., Zimmer, A., Barber, R.D., Cann, I., Graham, D.E., Grahame, D.A., Guss, A.M., Hedderich, R., Ingram-Smith, C., Kuettner, H.C., Krzycki, J.A., Leigh, J.A., Li, W., Liu, J., Mukhopadhyay, B., Reeve, J.N., Smith, K., Springer, T.A., Umayam, L.A., White, O., White, R.H., Conway de Macario, E., Ferry, J.G., Jarrell, K.F., Jing, H., Macario, A.J., Paulsen, I., Pritchett, M., Sowers, K.R., Swanson, R.V., Zinder, S.H., Lander, E., Metcalf, W.W. and Birren, B. (2002) The genome of *M. acetivorans* reveals extensive metabolic and physiological diversity, *Genome Res* **12**, 532-42

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

56 Hendrickson, E.L., Kaul, R., Zhou, Y., Bovee, D., Chapman, P., Chung, J., Conway de Macario, E., Dodsworth, J., Gillett, W., Graham, D.E., Haydock, A.K., Kang, A., Land, M.L., Levy, R., Lie, T.J., Major, T., Moore, B., Porat, I., Overbeek, R., Palmeiri, A., Rouse, G., Saenphimmachak, C., Soll, D., Whitman, W.B., Larimer, F.W., Olson, M.V. and Leigh, J.A. (2004) Complete genome sequence of the mesophilic hydrogenotrophic methanogen *Methanococcus maripaludis*, *Unpublished*

57 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., Tatusov, R.L., Wolf, Y.I., Stetter, K.O., Malykh, A.G., Koonin, E.V. and Kozyavkin, S.A. (2002) The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens, *Proc Natl Acad Sci U S A* **99**, 4644-9

58 Smith, D.R., Doucette-Stamm, L.A., Deloughery, C., Lee, H., Dubois, J., Aldredge, T., Bashirzadeh, R., Blakely, D., Cook, R., Gilbert, K., Harrison, D., Hoang, L., Keagle, P., Lumm, W., Pothier, B., Qiu, D., Spadafora, R., Vicaire, R., Wang, Y., Wierzbowski, J., Gibson, R., Jiwani, N., Caruso, A., Bush, D., Reeve, J.N. and et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* deltaH: functional analysis and comparative genomics, *J Bacteriol* **179**, 7135-55

59 Cohen, G.N., Barbe, V., Flament, D., Galperin, M., Heilig, R., Lecompte, O., Poch, O., Prieur, D., Querellou, J., Ripp, R., Thierry, J.C., Van der Oost, J., Weissenbach, J., Zivanovic, Y. and Forterre, P. (2003) An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*, *Mol Microbiol* **47**, 1495-512

60 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**, 134-57

61 Kawarabayasi, Y., Sawada, M., Horikawa, H., Haikawa, Y., Hino, Y., Yamamoto, S., Sekine, M., Baba, S., Kosugi, H., Hosoyama, A., Nagai, Y., Sakai, M., Ogura, K., Otsuka, R., Nakazawa, H., Takamiya, M., Ohfuku, Y., Funahashi, T., Tanaka, T., Kudoh, Y., Yamazaki, J., Kushida, N., Oguchi, A., Aoki, K. and Kikuchi, H. (1998) Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3, *DNA Res* **5**, 55-76

62 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**, 352-63

63 Futterer, O., Angelov, A., Liesegang, H., Gottschalk, G., Schleper, C., Schepers, B., Dock, C., Antranikian, G. and Liebl, W. (2004) Genome sequence of *Picrophilus torridus* and its implications for life around pH 0, *Proc Natl Acad Sci U S A* **101**, 9091-6

64 Ruepp, A., Graml, W., Santos-Martinez, M.L., Koretke, K.K., Volker, C., Mewes, H.W., Frishman, D., Stocker, S., Lupas, A.N. and Baumeister, W. (2000) The genome sequence of the thermoacidophilic scavenger *Thermoplasma acidophilum*, *Nature* **407**, 508-13

65 Kawashima, T., Amano, N., Koike, H., Makino, S., Higuchi, S., Kawashima-Ohya, Y., Watanabe, K., Yamazaki, M., Kanehori, K., Kawamoto, T., Nunoshiba, T., Yamamoto, Y., Aramaki, H., Makino, K. and Suzuki, M. (2000) Archaeal adaptation to higher temperatures revealed by genomic sequence of *Thermoplasma volcanium*, *Proc Natl Acad Sci U S A* **97**, 14257-62

66 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**, 276-8

67 Snel, B., Bork, P. and Huynen, M.A. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content, *Genome Res* **12**, 17-25

68 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**, 485-7

69 Aravind, L., Watanabe, H., Lipman, D.J. and Koonin, E.V. (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes, *Proc Natl Acad Sci U S A* **97**, 11319-24

70 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**, 684-93

71 Maeder, D.L., Weiss, R.B., Dunn, D.M., Cherry, J.L., Gonzalez, J.M., DiRuggiero, J. and Robb, F.T. (1999) Divergence of the hyperthermophilic archaea Pyrococcus furiosus and P. horikoshii inferred from complete genomic sequences, *Genetics* **152**, 1299-305

72 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**, 608-28

73 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool, *J Mol Biol* **215**, 403-10

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

75 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**, 3389-402

76 Eddy, S.R. (1995) Multiple alignment using hidden Markov models, *Proc Int Conf Intell Syst Mol Biol* **3**, 114-20

77 Ponting, C.P., Schultz, J., Copley, R.R., Andrade, M.A. and Bork, P. (2000) Evolution of domain families, *Adv Protein Chem* **54**, 185-244

78 Jacob, F. (2001) Complexity and tinkering, *Ann N Y Acad Sci* **929**, 71-3

79 Jacob, F. (1977) Evolution and tinkering, *Science* **196**, 1161-6

80 Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. and Bork, P. (2004) SMART 4.0: towards genomic data integration, *Nucleic Acids Res* **32 Database issue**, D142-4

81 Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R. (2004) The Pfam protein families database, *Nucleic Acids Res* **32 Database issue**, D138-41

82 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., Liebert, C.A., Liu, C., Madej, T., Marchler, G.H., Mazumder, R., Nikolskaya, A.N., Panchenko, A.R., Rao, B.S., Shoemaker, B.A., Simonyan, V., Song, J.S., Thiessen, P.A., Vasudevan, S., Wang, Y., Yamashita, R.A., Yin, J.J. and Bryant, S.H. (2003) CDD: a curated Entrez database of conserved domain alignments, *Nucleic Acids Res* **31**, 383-7

83 Hulo, N., Sigrist, C.J., Le Saux, V., Langendijk-Genevaux, P.S., Bordoli, L., Gattiker, A., De Castro, E., Bucher, P. and Bairoch, A. (2004) Recent improvements to the PROSITE database, *Nucleic Acids Res* **32 Database issue**, D134-7

84 Fitch, W.M. (1970) Distinguishing homologous from analogous proteins, *Syst Zool* **19**, 99-113

85 Sonnhammer, E.L. and Koonin, E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes, *Trends Genet* **18**, 619-20

86 Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families, *Science* **278**, 631-7

87 Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S. and Koonin, E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context, *Genome Res* **11**, 356-72

88 Huynen, M.A. and Snel, B. (2000) Gene and context: integrative approaches to genome analysis, *Adv Protein Chem* **54**, 345-79

89 Jacob, F. and Monod, J. (1961) Genetic regulatory mechanisms in the synthesis of proteins, *J Mol Biol* **3**, 318-56

90 Huynen, M.A. and Bork, P. (1998) Measuring genome evolution, *Proc Natl Acad Sci U S A* **95**, 5849-56

91 Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact, *Trends Biochem Sci* **23**, 324-8

92 Huynen, M., Snel, B., Lathe, W. and Bork, P. (2000) Exploitation of gene context, *Curr Opin Struct Biol* **10**, 366-70

93 Huynen, M., Snel, B., Lathe, W., 3rd and Bork, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences, *Genome Res* **10**, 1204-10

94 Galperin, M.Y. and Koonin, E.V. (2000) Who's your neighbor? New computational approaches for functional genomics, *Nat Biotechnol* **18**, 609-13

95 Korbel, J.O., Jensen, L.J., von Mering, C. and Bork, P. (2004) Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs, *Nat Biotechnol* **22**, 911-7

96 Beck, C.F. and Warren, R.A. (1988) Divergent promoters, a common form of gene organization, *Microbiol Rev* **52**, 318-26

97 Koonin, E.V., Mushegian, A.R. and Bork, P. (1996) Non-orthologous gene displacement, *Trends Genet* **12**, 334-6

98 Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences, *Science* **285**, 751-3

99 Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events, *Nature* **402**, 86-90

100 Snel, B., Bork, P. and Huynen, M. (2000) Genome evolution. Gene fusion versus gene fission, *Trends Genet* **16**, 9-11

101 Ibba, M. and Soll, D. (2004) Aminoacyl-tRNAs: setting the limits of the genetic code, *Genes Dev* **18**, 731-8

102 Randau, L., Munch, R., Hohn, M.J., Jahn, D. and Soll, D. (2005) Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5'- and 3'-halves, *Nature* **433**, 537-41

103 Tumbula, D., Vothknecht, U.C., Kim, H.S., Ibba, M., Min, B., Li, T., Pelaschier, J., Stathopoulos, C., Becker, H. and Soll, D. (1999) Archaeal aminoacyl-tRNA synthesis: diversity replaces dogma, *Genetics* **152**, 1269-76

104 Randau, L., Pearson, M. and Soll, D. (2005) The complete set of tRNA species in Nanoarchaeum equitans, *FEBS Lett* **579**, 2945-7

105 Makarova, K.S., Aravind, L., Grishin, N.V., Rogozin, I.B. and Koonin, E.V. (2002) A DNA repair system specific for thermophilic Archaea and bacteria predicted by genomic context analysis, *Nucleic Acids Res* **30**, 482-96

106 Guy, C.P., Majernik, A.I., Chong, J.P. and Bolt, E.L. (2004) A novel nuclease-ATPase (Nar71) from archaea is part of a proposed thermophilic DNA repair system, *Nucleic Acids Res* **32**, 6176-86

107 Ettema, T.J.G., Brinkman, A.B., Lamers, P.P., Kornet, N.G., de Vos, W.M. and van der Oost, J. (2004) An archaeal copper-responsive gene cluster is controlled by a TRASH-domain containing regulator, CopT, *unpublished*

108 Ettema, T.J., Huynen, M.A., de Vos, W.M. and van der Oost, J. (2003) TRASH: a novel metal-binding domain predicted to be involved in heavy-metal sensing, trafficking and resistance, *Trends Biochem Sci* **28**, 170-3

109 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**, 529-40

110 Daugherty, M., Vonstein, V., Overbeek, R. and Osterman, A. (2001) Archaeal shikimate kinase, a new member of the GHMP-kinase family, *J Bacteriol* **183**, 292-300

111 Makarova, K.S. and Koonin, E.V. (2003) Filling a gap in the central metabolism of archaea: prediction of a novel aconitase by comparative-genomic analysis, *FEMS Microbiol Lett* **227**, 17-23

112 Galperin, M.Y., Aravind, L. and Koonin, E.V. (2000) Aldolases of the DhnA family: a possible solution to the problem of pentose and hexose biosynthesis in archaea, *FEMS Microbiol Lett* **183**, 259-64

113 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**, 28710-8

114 van der Oost, J., Huynen, M.A. and Verhees, C.H. (2002) Molecular characterization of phosphoglycerate mutase in archaea, *FEMS Microbiol Lett* **212**, 111-20

115 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**, 7754-62

116 Morett, E., Korbel, J.O., Rajan, E., Saab-Rincon, G., Olvera, L., Olvera, M., Schmidt, S., Snel, B. and Bork, P. (2003) Systematic discovery of analogous enzymes in thiamin biosynthesis, *Nat Biotechnol* **21**, 790-5

117 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**, 48949-59

118 Gabaldon, T. and Huynen, M.A. (2004) Prediction of protein function and pathways in the genome era, *Cell Mol Life Sci* **61**, 930-44

119 Bisbal, C., Martinand, C., Silhol, M., Lebleu, B. and Salehzada, T. (1995) Cloning and characterization of a RNAse L inhibitor. A new component of the interferon-regulated 2-5A pathway, *J Biol Chem* **270**, 13308-17

120 Zimmerman, C., Klein, K.C., Kiser, P.K., Singh, A.R., Firestein, B.L., Riba, S.C. and Lingappa, J.R. (2002) Identification of a host protein essential for assembly of immature HIV-1 capsids, *Nature* **415**, 88-92

121 Estevez, A.M., Haile, S., Steinbuchel, M., Quijada, L. and Clayton, C. (2004) Effects of depletion and overexpression of the Trypanosoma brucei ribonuclease L inhibitor homologue, *Mol Biochem Parasitol* **133**, 137-41

122 Kispal, G., Sipos, K., Lange, H., Fekete, Z., Bedekovics, T., Janaky, T., Bassler, J., Aguilar Netz, D.J., Balk, J., Rotte, C. and Lill, R. (2005) Biogenesis of cytosolic ribosomes requires the essential iron-sulphur protein Rli1p and mitochondria, *Embo J* **24**, 589-98

123 Yarunin, A., Panse, V.G., Petfalski, E., Dez, C., Tollervey, D. and Hurt, E.C. (2005) Functional link between ribosome formation and biogenesis of iron-sulfur proteins, *Embo J* **24**, 580-8

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**, 407-14

125 Velculescu, V.E., Zhang, L., Vogelstein, B. and Kinzler, K.W. (1995) Serial analysis of gene expression, *Science* **270**, 484-7

126 Chien, C.T., Bartel, P.L., Sternglanz, R. and Fields, S. (1991) The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest, *Proc Natl Acad Sci U S A* **88**, 9578-82

127 Covert, M.W., Knight, E.M., Reed, J.L., Herrgard, M.J. and Palsson, B.O. (2004) Integrating high-throughput and computational data elucidates bacterial networks, *Nature* **429**, 92-6

128 Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A. and Gerstein, M. (2004) Genomic analysis of regulatory network dynamics reveals large topological changes, *Nature* **431**, 308-12

129 Hood, L. and Galas, D. (2003) The digital code of DNA, *Nature* **421**, 444-8

130 Baliga, N.S., Bjork, S.J., Bonneau, R., Pan, M., Iloanusi, C., Kottemann, M.C., Hood, L. and DiRuggiero, J. (2004) Systems level insights into the stress response to UV radiation in the halophilic archaeon Halobacterium NRC-1, *Genome Res* **14**, 1025-35

Chapter3

Modularity in the gain and loss of genes: applications for function prediction

Thijs J.G. Ettema
John van der Oost
Martijn A. Huynen

Genes that are clustered on multiple genomes and are likely to functionally interact tend to be gained or lost together during genome evolution. Here, we demonstrate that exceptions to this pattern indicate relatively distant functional interactions between the encoded proteins. Hence, this can be used to divide predicted clusters of functionally interacting proteins into sub-clusters, and as such, to refine the prediction of their function and functional interactions.

Trends in Genetics 17(9), 485-487 (2001)

INTRODUCTION

The increasing availability of sequenced genomes allows for the analysis of differences between genomes in terms of the gain and loss of genes. This enables the study of modularity in genome evolution: do genes that are functionally linked, for example because they encode enzymes from the same pathway, tend to be gained and lost simultaneously? The modularity of genome evolution bears relevance for the prediction of gene function, as the co-occurrence of genes in genomes has been proposed[1] and demonstrated[2] to indicate functional relations between their protein products. Recently, modularity has been observed in the loss of genes in one species: *Saccharomyces cerevisiae*[3]. Here, we present the first systematic analysis of the gain and loss of genes and their modularity within the first genus of which three genomes are available, *Pyrococcus furiosus*, *Pyrococcus abyssi* and *Pyrococcus horikoshii*.

FUNCTIONAL PATTERNS IN THE GAIN AND LOSS OF GENES

In the three pyrococcal genomes, a total of 1071 genes are present in at least one *Pyrococcus* species that have no orthologs in the other sequenced Archaea. Thus, this subset of genes has most likely been gained in the evolution of *Pyrococcus*. Conversely, 325 genes are present in both the Euryarchaea and the Crenarchaea, but absent in at least one *Pyrococcus* species, and so have probably been lost. A functional classification of the gained and lost genes along the cluster of orthologous genes (COG) scheme[4] revealed the dominance of a few classes. Genes involved in ‘amino acid transport and metabolism’ and ‘energy production and conversion’ have both been gained (12% and 15%, respectively) and lost more frequently (22% for both) than other functional classes. Interestingly, the gained genes that are involved in ‘amino acid transport and metabolism’ and those of a third functional class of genes that showed significant gain, that of ‘carbohydrate transport and metabolism’ (13%), are dominated by heterotrophic functions, indicating the evolution of *Pyrococcus* towards a heterotrophic lifestyle. The corresponding proteins are involved in the import and catabolism of amino acids and carbohydrates, including a putative galactoside utilization cluster (Figure 3.1b), α-amylases, cellulases and several other hydrolases capable of degrading α- and β-linked carbohydrate substrates (http://www.dove.embl-heidelberg.de/Pyrococcus).

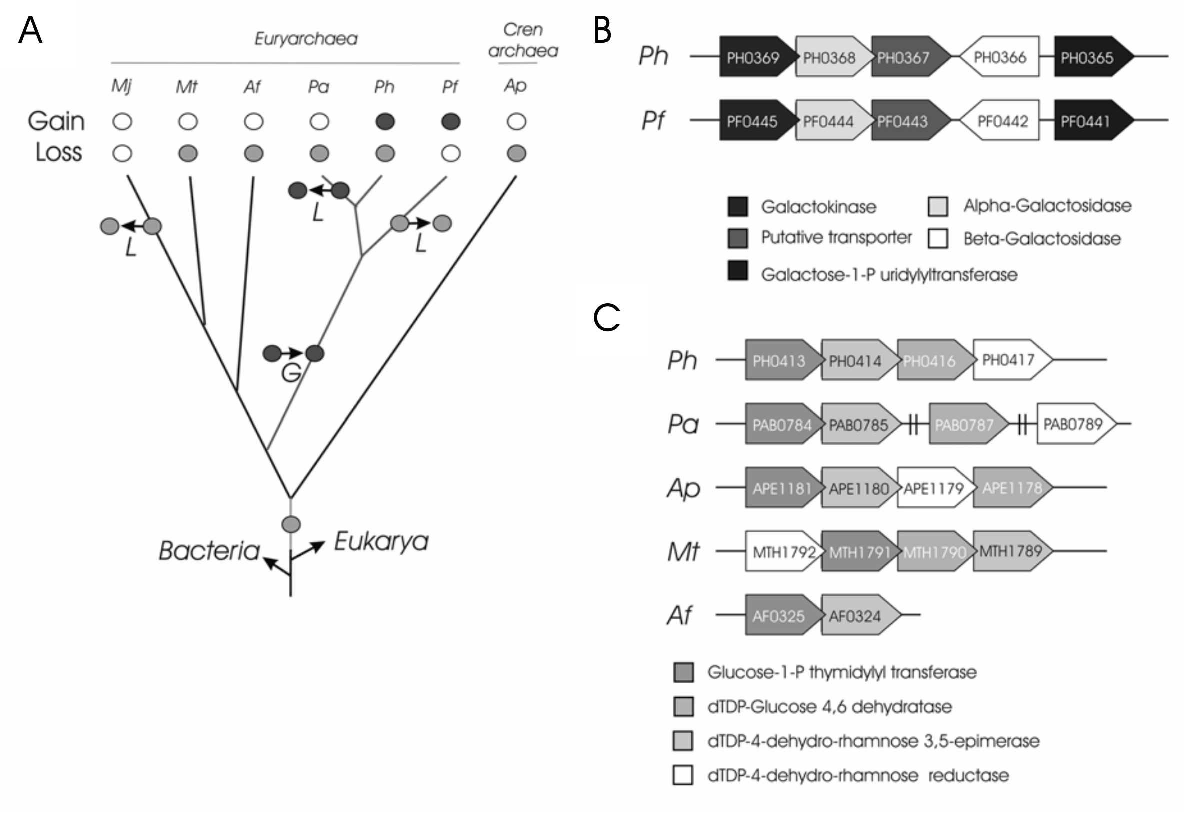


Figure 3.1 (NB: For COLOR version of this figure, see APPEDIX I) Gain and loss of genes in *Pyrococcus*. **(a)** The prediction of gains and losses of genes was based on their phylogenetic distribution, using a genome-based phylogenetic tree[5]. A pyrococcal gain [purple circles, corresponding with cluster in (b)] was defined as a gene that is present in at least one of the *Pyrococci* but has no detectable orthologous counterparts in the other Archaea, thereby selecting genes that were potentially horizontally transferred, duplicated or invented in the pyrococcal branch (*red*). A pyrococcal loss was defined as a gene that has an ortholog in *Aeropyrum pernix* and in at least one of the sequenced Euryarchaea (Af, Mt, Mj), but has no ortholog in one or more *Pyrococci*. *Green* circles correspond to the evolutionary scenario for the example described in (c). Using these conditions, genes were selected that were probably present in the last common archaeal ancestor (*green* branch) and were lost in the pyrococcal branch (*red*). Whether genes were shared between archaeal taxa was determined using an operational definition of orthology. Genes were orthologous when Smith–Waterman-based sequence comparisons[6] between pairs of genes of two genomes displayed highest, significant (e <0.01) bi-directional best hits including the possibility of fission and fusion[1]. **(b)** The presence of a galactoside utilization cluster in both *P. furiosus* and *P. horikoshii* and the absence of these genes in the other Archaea is an example of a gain of functionally interacting genes in these *Pyrococci*. Notice that the *Pyrococci* themselves are phylogenetically too close to consider conservation of gene order among them a significant indication for functional interaction; the galactokinase and the galactose-1-phosphate uridylyl transferase are clustered in numerous Bacteria. **(c)** The presence of a polymeric sugar biosynthesis gene cluster that is probably involved in cell wall biosynthesis in *A. pernix*, *P. abyssi*, *P. horikoshii* and the other Euryarchaeota, but not in *P. furiosus* is an example of a loss event of functionally interacting genes in *P. furiosus*. Notice that the gene cluster is partly lost in *A. fulgidus*. Clusters of functionally interacting genes were

Figure 3.1 (continued...)
detected using STRING (http://www.bork.EMBL-Heidelberg.DE/STRING/). Abbreviations: L, loss event; G, gain event; Mt, *Methanobacterium thermoautotrophicum*; Mj, *Methanococcus jannaschii*; Af, *Archaeoglobus fulgidus*; Ph, *Pyrococcus horikoshii*; Pa, *Pyrococcus abyssi*; Pf, *Pyrococcus furiosus*; Ap, *Aeropyrum pernix*. For availability of genomic coding sequences see http://www.TIGR.org, except for *P. furiosus*[7] (see http://comb5156.umbi.umd.edu/genemate/). To prevent that genes that are annotation artefacts are counted as gains here we only count pyrococcal genes for which we could detect at least one homolog or that are at least 150 amino acids long. Furthermore, all genes that were detected in only a subset of the *Pyrococci* were searched for in the DNA of the other *Pyrococci* using TBLASTN[8], and if found, added to the predicted genes in that genome. For an additional comparison of the pyrococcal genomes, see[9].

Functional coupling within the gain and loss of genes

The fact that the genes from a few functional classes are most frequently gained or lost suggests that their functions are coupled. To obtain an increased level of resolution of the functional coupling between genes that are gained or lost, we examined how often genes are co-gained or co-lost with other genes that they tend to be clustered with on the genome. The occurrence of genes in the same neighbourhood on multiple, phylogenetically distant genomes is a strong indication of functional interactions between their proteins: it tends to reflect a physical interaction between the corresponding proteins or an involvement in the same metabolic pathway[10-12]. The conserved gene neighbourhood detection tool STRING[13] was used to predict functional links for the genes that were gained or lost. Functional links could thus be detected for 163 genes that were gained, and for 177 genes that were lost. Both the gain and the loss of genes showed a significant degree of modularity. In 76 cases (47%), genes were gained together with genes with which they tend to occur in operons. For the losses, the percentage is slightly but significantly higher, here we found 104 cases (59%, $P < 0.05$, χ^2) where genes were co-lost with genes that occurred in the same operon.

Most of the functionally coupled gains consist of uptake systems, such as ABC-transporters (including a maltose transporter), or complete utilization clusters, such as a galactoside utilization cluster (Figure 3.1b). Another example is a cluster of genes in *P. abyssi* and *P. furiosus* that is probably involved in glycerol processing. It consists of a glycerophosphoryl diester phosphodiesterase (PAB0180), a glycerol kinase (PAB2406), a glycerol-3-phosphate dehydrogenase (PAB0183), a putative oxidoreductase with unknown specificity (PAB0184) and a hypothetical protein (PAB0185).

The largest number of lineage specific gene losses[3] was observed in *P. horikoshii* (37 cases, compared with 10 and 3 for *P. furiosus* and *P. abyssi*, respectively). Aside from massive gene loss in the aromatic amino acid biosynthesis operons[14], *P. horikoshii* has lost a phosphate ABC transporter that is present in most sequenced archaeal and bacterial genomes, except in a few parasitic organisms (e.g. Chlamydia species and *Rickettsia prowazekii*). No obvious candidate for an alternative phosphate uptake system could be found in the *P. horikoshii* genome. For a complete overview of the modular gains and losses see http://www.dove.embl-heidelberg.de/Pyrococcus.

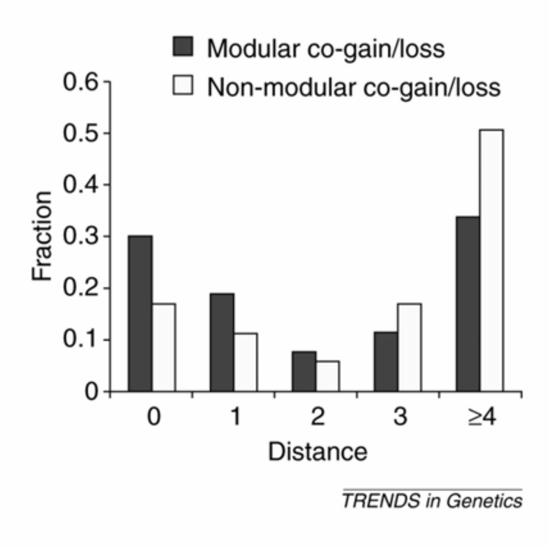


Figure 3.2 Distribution of metabolic distances between enzymes that occur in the same neighbourhood on at least two, phylogenetically distant genomes, and that have been co-gained/co-lost (*grey*) or not co-gained/co-lost (*white*) in the evolution of the *Pyrococcus* genus. Genes that tend to be clustered and that have been gained or lost together have a more direct functional interaction than genes that have been gained or lost independently: as the metabolic distance increases, the relative fraction of co-gains/co-losses compared with the not co-gained/co-lost decreases. Fractions were calculated from the total number of co-gained/co-lost and not co-gained/co-lost genes, respectively. The metabolic distance between two enzymes is defined as the number of metabolites that separates them in the Kyoto Encyclopedia of Genes and Genomes (KEGG)[15] metabolic maps. This distance is zero if the enzymes are involved in the same conversion, one if they are involved in subsequent conversions, and so on. Distances were derived from the BRITE database (http://www.genome.ad.jp/brite/). When a gene tends to be clustered with multiple other genes, the median of the metabolic distances with these other genes was taken.

Applications for the prediction of functional interactions between proteins

The most interesting aspect of these results is not the overlap between genes that have been gained or lost together and genes that tend to be clustered on genomes, but rather the discrepancy: genes that tend to be clustered but that have not been lost or gained together. An example is the four-gene cluster encoding the biosynthesis of a polymeric sugar (Figure 3.1c) that is completely lost in *P. furiosus*, but only partly lost in *Archaeoglobus* *fulgidus*. The genes that are lost in *A. fulgidus* encode the enzymes for the last two steps of the pathway. Another example is the *arg*H and *arg*G genes that have been lost in *P. abyssi* and *P. horikoshii*, even though the genes with which they tend to be clustered (*arg*BCDEF/IJ) have been retained. As in the first example, the functions of ArgH and ArgG (arginosuccinate lyase and synthase respectively; both involved in the urea cycle) are relatively peripheral to those of the other gene products that have been retained.

The above examples suggest that genes that tend to be clustered and that have been gained or lost together have a more direct functional interaction than genes that have been gained or lost independently. A systematic, quantitative analysis (Figure 3.2) confirms this idea: the fraction of gene products that are involved in the same metabolic reaction, or subsequent reactions, is higher when the genes have been gained or lost together than when they have been gained or lost independently.

An example of how this information can be used for function prediction is the *Escherichia coli* gene b2748 (COG2919), whose ortholog in *Bacillus subtilis* has been implicated in septum formation[16], but whose molecular function has not been elucidated. Even though this gene occurs on four genomes as neighbour of the gene encoding 4-diphosphocytidyl-2-methyl-d-erythritol synthase (isoprenoid synthesis), the isoprenoid synthesis gene has been gained by *P. horikoshii*, but the septum formation gene has not. It is very unlikely that the genes have a very direct functional interaction. Conversely, the membrane proteins encoding genes that have been gained by the *Pyrococci* (PH1801 and PH1802) always occur together as neighbours on prokaryotic genomes, although with a different local gene context. Thus, these genes are likely to have a very direct relationship, possibly in the form of a physical interaction between their products. For a complete list of genes that tend to be clustered but have not been gained or lost together see http://www.dove.embl-heidelberg.de/Pyrococcus.

By analyzing whether similar phyletic distributions of genes are the result of co-gain and co-loss rather than vertical inheritance, we can estimate the number of independent co-occurrences of genes, which is a requirement for a statistical verification of such analysis. Still, the numbers of gains or losses that can be observed in individual lineages are generally too large to predict functional interactions from the co-gain or co-loss of genes alone. Combining various types of genomic association for the prediction of functional interaction reduces the fraction of false positives[17]. Conservation of gene order does produce, however, very few false positives[11]. Here, combining the information that genes have been lost or gained together can make the functional relations that can be predicted more direct. Thus, by combining conserved gene context with the co-occurrence of genes in genomes, we can divide the large sets of functionally interacting proteins that are predicted by the first method into subsets. Such division is a necessary requirement to refine predictions of functional interaction of proteins based on genomic association data to a higher resolution and of use for experimental verification.

ACKNOWLEDGEMENT

This research was supported by BMBF. We thank Steffen Schmidt for the calculation of the metabolic differences between enzymes.

REFERENCES

1 Huynen, M.A. and Bork, P. (1998) Measuring genome evolution, *Proc Natl Acad Sci U S A* **95**, 5849-56

2 Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles, *Proc Natl Acad Sci U S A* **96**, 4285-8

3 Aravind, L., Watanabe, H., Lipman, D.J. and Koonin, E.V. (2000) Lineage-specific loss and divergence of functionally linked genes in eukaryotes, *Proc Natl Acad Sci U S A* **97**, 11319-24

4 Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families, *Science* **278**, 631-7

5 Snel, B., Bork, P. and Huynen, M.A. (1999) Genome phylogeny based on gene content, *Nat Genet* **21**, 108-10

6 Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences, *J Mol Biol* **147**, 195-7

7 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**, 134-57

8 Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool, *J Mol Biol* **215**, 403-10

9 Lecompte, O., Ripp, R., Puzos-Barbe, V., Duprat, S., Heilig, R., Dietrich, J., Thierry, J.C. and Poch, O. (2001) Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea, *Genome Res* **11**, 981-93

10 Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact, *Trends Biochem Sci* **23**, 324-8

11 Huynen, M., Snel, B., Lathe, W., 3rd and Bork, P. (2000) Predicting protein function by genomic context: quantitative evaluation and qualitative inferences, *Genome Res* **10**, 1204-10

12 Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D. and Maltsev, N. (1999) The use of gene clusters to infer functional coupling, *Proc Natl Acad Sci U S A* **96**, 2896-901

13 Snel, B., Lehmann, G., Bork, P. and Huynen, M.A. (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene, *Nucleic Acids Res* **28**, 3442-4

14 Maeder, D.L., Weiss, R.B., Dunn, D.M., Cherry, J.L., Gonzalez, J.M., DiRuggiero, J. and Robb, F.T. (1999) Divergence of the hyperthermophilic archaea *Pyrococcus furiosus* and *P. horikoshii* inferred from complete genomic sequences, *Genetics* **152**, 1299-305

15 Kanehisa, M. and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes, *Nucleic Acids Res* **28**, 27-30

16 Levin, P.A. and Losick, R. (1994) Characterization of a cell division gene from *Bacillus subtilis* that is required for vegetative and sporulation septum formation, *J Bacteriol* **176**, 1451-9

17 Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function, *Nature* **402**, 83-6

Chapter 4

TRASH: a novel metal-binding domain predicted to be involved in heavy-metal sensing, trafficking and resistance

Thijs J. G. Ettema
Martijn A. Huynen
Willem M. de Vos
John van der Oost

We describe a previously undetected domain – TRASH – containing a well-conserved cysteine motif that we anticipate to be involved in metal coordination. TRASH is encoded by multiple prokaryotic genomes and is present in transcriptional regulators, cation-transporting ATPases and hydrogenases, and is also present as a stand-alone module. The observed domain associations and conserved genome context of TRASH-encoding genes in prokaryotic genomes suggest that TRASH constitutes a novel component in metal trafficking and heavy-metal resistance. The role of the multiple copies of TRASH that are present in vertebrate proteins remains to be elucidated.

Trends in Biochemical Sciences 28(4), 170-173 (2003)

INTRODUCTION

Metals play an essential role as trace elements in many biological systems; current estimates indicate that over half of all proteins are metalloproteins, containing metal ions either as a structural component or as a catalytic cofactor[1]. Both eukaryotes and prokaryotes have evolved several mechanisms that ensure efficient metal homeostasis, such as sequestering of metals by metallothioneins[2,3], the efflux of metals by P-type cation-transporting ATPases (CTAs)[4] and trafficking of metal ions by metallochaperones[5,6]. Metallochaperones have been shown to interact with metal ions and deliver them to several targets in the cell, the best example being the transfer of copper between the human copper chaperone HAH1p and the copper-transporting ATPases that are associated with Wilson and Menkes disease[7-10]. Here, we present a previously undetected metallochaperone-like domain, which is encoded by multiple bacterial and archaeal genomes and contains distant homologs in higher eukaryotes.

RESULTS AND DISCUSSION

Characterization of the TRASH domain

In the course of analyzing heavy-metal resistance in prokaryotes, we found a distinct clustering of genes that are involved in copper homeostasis in several phylogenetically distinct genomes (Figure 4.1a). These include a homolog of the Menkes and Wilson disease-associated copper-transporting ATPase (e.g. Ta1143), an archaea-specific transcriptional regulator (e.g. SSO2652) and a small open reading frame (ORF) with unknown function (e.g. Ta1144). All three proteins contain a distinct cysteine signature, Cys-Xaa19–22-Cys-Xaa3-Cys (CxCxC); it should be noted that transcriptional regulators contain the extended cysteine motif C-Xaa2-Cys-Xaa19–22-Cys-Xaa3-Cys (CxCxCxC).

To detect functional relatives of this domain, several PSI–BLAST searches[11] were conducted against the non-redundant database at the National Center for Biotechnology Information (NCBI; http://www.ncbi.nih.gov) using the regions of proteins containing the cysteine motif. For example, when initiating a search (BLOSUM80, inclusion threshold of 0.005) with the small ORF of *Thermoplasma acidophilum* (Ta1144, residues 3–40), several hits with N-terminal regions of other CTAs (e.g. mlr5325, 2e-6) and numerous other small, uncharacterized Ta1144-like proteins (e.g. STS190, 1e-8) were detected within the first iteration. In addition, hits were obtained with the N-terminal region of the β-subunit of coenzyme F420-dependent hydrogenases (e.g. MM1225 in iteration 2, 4e-8) that are encoded by all three sequenced *Methanosarcina* genomes (Figure 4.1b). The search converged after four iterations and retrieved >40 instances of the novel domain.

Because it was anticipated that Ta1144-like proteins could have been easily overlooked in the genome annotation process owing to their small size (generally <70 residues), we screened the finished genomes for the presence of Ta1144 homologs using tBLASTn. Indeed, this search retrieved several previously undetected genes encoding Ta1144-like proteins, some of which

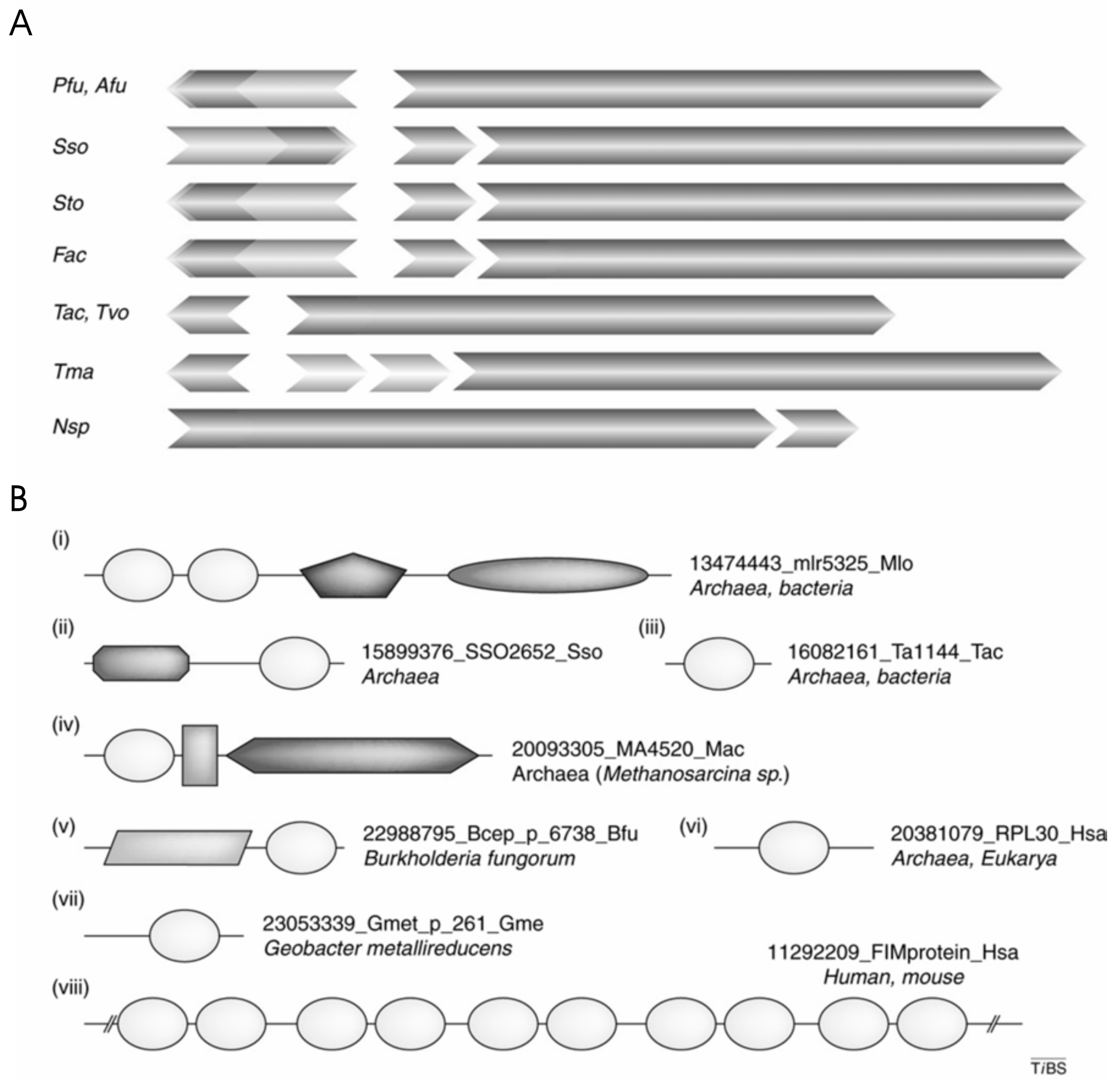
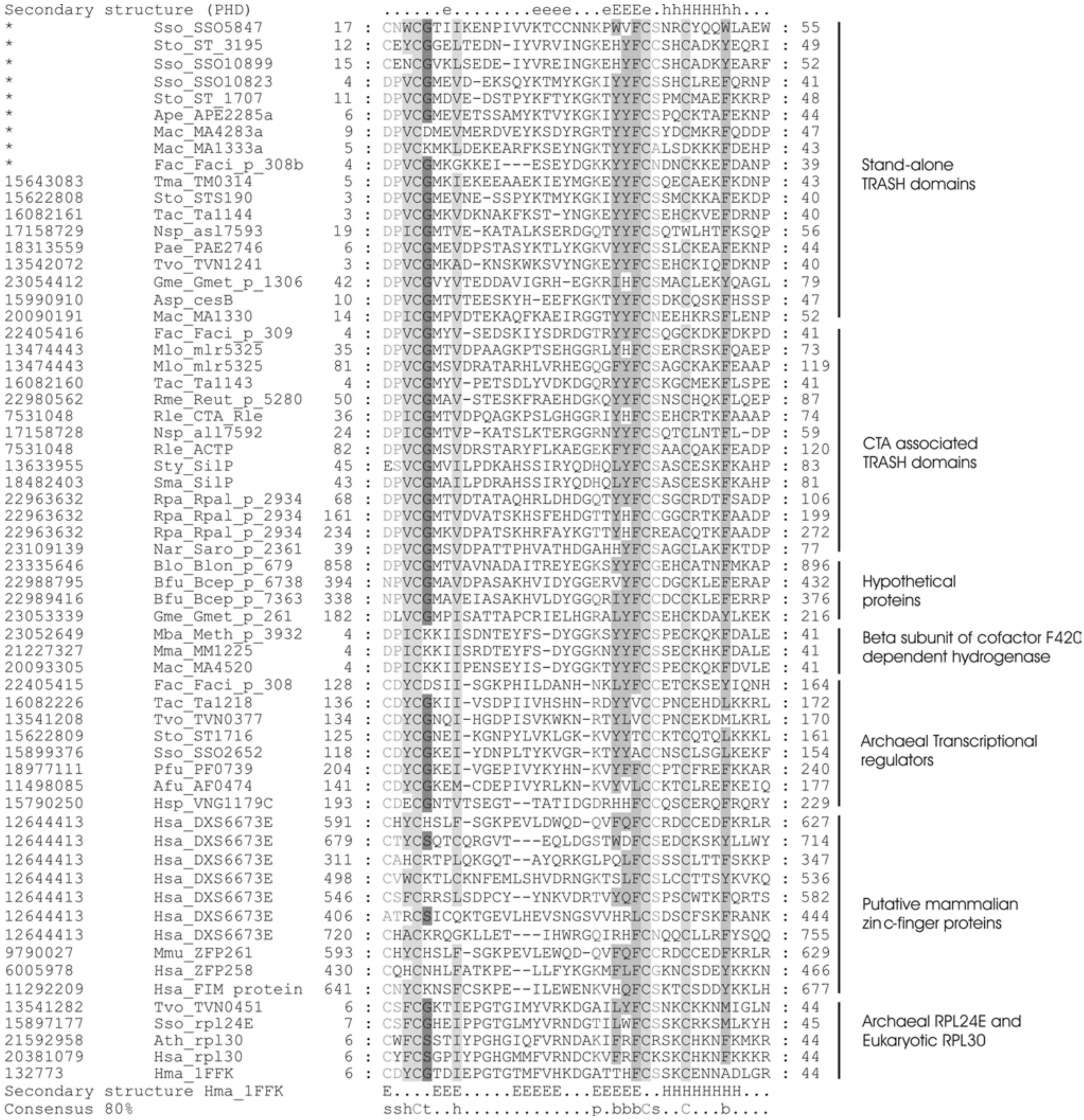


Figure 4.1 (previous page; NB: For COLOR version of this figure, see APPEDIX I) Physical associations of the TRASH domain. **(a)** Operon and divergon structure of genes encoding TRASH-containing copper chaperones, copper-transporting ATPases (*dark blue*) and transcriptional regulators (*light blue*) are shown. Genes encoding stand-alone TRASH domains (copper chaperones) are indicated in *pink*, TRASH domains of copper-transporting ATPases and transcriptional regulators are indicated in *green*. Unrelated genes are *gray*, arrows are not drawn to scale. The following genes are shown: Pfu: PF0739, PF0740; Afu: AF0474, AF0473; Sso: SSO2652, SSO10823*, SSO2651; Sto: ST1716, STS190, ST1715; Fac: Faci_p_308, Faci_p_308b*, Faci_p_309; Tac: Ta1144, Ta1143; Tvo: TVN6226, TVN6221; Tma: TM0314, TM0317; Nsp: asl7592, asl7593. Genes indicated with an asterisk were ignored during genome annotation and are available at http://www.ftns.wau.nl/micr/bacgen/TRASH/stand-alone.htm. **(b)** Domain architectures of TRASH-containing proteins. Proteins are approximately drawn to scale and denoted by their GenBank Identifier, followed by the species abbreviation and gene name. In addition, the phyletic distribution of the proteins is displayed in italics. Domain architectures were analyzed using SMART[12] and Pfam[13]. The TRASH-containing proteins are subdivided into the following classes: (*i*) P-type cation-transporting ATPases, typically containing one to three N-terminally fused TRASH domains (*yellow*), an E1-E2_ATPase domain (PF00122; *pink*) and a hydrolase domain (PF00702; *light blue*); (*ii*) transcriptional regulators containing an N-terminal helix–turn–helix motif (HTH_ASNC; *light green*) and a C-terminal TRASH domain; (*iii*) metallochaperone-like proteins, consisting of a single TRASH domain only; (*iv*) β-subunit of coenzyme F420-dependent hydrogenase, containing an N-terminal TRASH domain, followed by a ferredoxin domain (Fer4; PF00037; *purple*) and the C-terminal F420 dependent hydrogenase catalytic domain (*dark blue*); (v) uncharacterized protein, containing an N-terminal domain (PF02625) that is related to xanthine and CO dehydrogenase maturation factor (COG1975) and a C-terminal TRASH domain; (*vi*) Archaeal (RPL24E) and eukaryote (RPL30) ribosomal proteins containing a single N-terminal TRASH domain. (*vii*) hypothetical protein containing a C-terminal TRASH domain; (*viii*) putative mammalian zinc-finger proteins, containing multiple tandem repeated TRASH domains. Abbreviations: Afu, *Archaeoglobus fulgidus*; Asp, *Acinetobacter* sp. BW3; Ath, *Arabodopsis thaliana*; Avi, *Azotobacter vinelandii*; Bfu, *Burkholderia fungorum*; Blo, *Bifidobacterium longum*; Fac, *Ferroplasma acidarmanus*; Gme, *Geobacter metallireducens*; Hma, *Haloarcula marismortui*; Hsa, *Homo sapiens*; Hsp, *Halobacterium* sp. NRC-1; Mac, *Methanosarcina acetivorans*; Mba, *Methanosarcina barkeri*; Mlo, *Mesorhizobium loti*; Mma, *Methanosarcina mazei*; Mmu, *Mus musculus*; Nar, *Novosphingobium aromaticivorans*; Nsp, *Nostoc* sp. PCC 7120; Pae, *Pyrobaculum aerophilum*; Pfu, *Pyrococcus furiosus*; Rle, *Rhizobium leguminosarum*; Rme, *Ralstonia metallidurans*; Rpa, *Rhodopseudomonas palustris*; Sma, *Serratia marcescens*; Sso, *Sulfolobus solfataricus*; Sto, *Sulfolobus tokodaii*; Sty, *Salmonella typhimurium*; Tac, *Thermoplasma acidophilum*; Tma, *Thermotoga maritima*; Tvo, *Thermoplasma volcanium*.

encoded the extended cysteine motif (CxCxCxC). Interestingly, when the latter candidates were included in an iterative Hidden Markov profile search (inclusion threshold 0.01) using HMMER2 (http://www.wustl.edu/), we were able to retrieve the aforementioned archaea-specific transcriptional regulators and, in addition, several vertebrate proteins that appeared to contain multiple copies of the cysteine signature. Further iterations of the HMMer search revealed distant sequence similarity between TRASH and both the archaeal ribosomal protein RPL24E and its eukaryote counterpart RPL30. Altogether, >60 proteins containing the CxCxC motif were retrieved during these searches. We decided to name the novel domain TRASH because of its anticipated involvement in trafficking, resistance and sensing of heavy metals.

A multiple alignment of the TRASH domains was constructed using ClustalW[14], which was manually adjusted using the PSI–BLAST pairwise alignments, and the secondary structure was predicted using PHD[15] (Figure 4.2). The sequence conservation of the TRASH domain is predominantly



```
Secondary structure (PHD)                ......e........eeee...eEEEe.hhHHHHhh...
*               Sso_SSO5847       17 : CNWCGTIIKENPIVVKTCCNNKPWVFCSNRCYQQWLAEW : 55
*               Sto_ST_3195       12 : CEYCGGELTEDN-IYVRVINGKEHYFCCSHCADKYEQRI : 49
*               Sso_SSO10899      15 : CENCGVKLSEDE-IYVREINGKEHYFCCSHCADKYEARF : 52
*               Sso_SSO10823       4 : DPVCGMEVD-EKSQYKTMYKGKIYYFCSSHCLREFQRNP : 41
*               Sto_ST_1707       11 : DPVCGMDVE-DSTPYKFTYKGKTYYFCSPMCMAEFKKRP : 48
*               Ape_APE2285a       6 : DPVCGMEVETSSAMYKTVYKGKIYYFCSPQCKTAFEKNP : 44
*               Mac_MA4283a        9 : DPVCDMEVMERDVEYKSDYRGRTYYFCSYDCMKRFQDDP : 47
*               Mac_MA1333a        5 : DPVCKMKLDEKEARFKSEYNGKTYYFCALSDKKKFDEHP : 43
*               Fac_Faci_p_308b    4 : DPVCGMKGKKEI---ESEYDGKKYYFCNDNCKKEFDANP : 39
15643083        Tma_TM0314         5 : DPVCGMKIEKEEAAEKIEYMGKEYYFCSQECAEKFKDNP : 43
15622808        Sto_STS190         3 : DPVCGMEVNE-SSPYKTMYKGKIYYFCSSMCKKAFEKDP : 40
16082161        Tac_Ta1144         3 : DPVCGMKVDKNAKFKST-YNGKEYYFCSEHCKVEFDRNP : 40
17158729        Nsp_asl7593       19 : DPICGMTVE-KATALKSERDGQTYYFCSQTWLHTFKSQP : 56
18313559        Pae_PAE2746        6 : DPVCGMEVDPSTASYKTLYKGKVYYFCSSLCKEAFEKNP : 44
13542072        Tvo_TVN1241        3 : DPVCGMKAD-KNSKWKSVYNGKEYYFCSEHCKIQFDKNP : 40
23054412        Gme_Gmet_p_1306   42 : DPVCGVYVTEDDAVIGRH-EGKRIHFCSMACLEKYQAGL : 79
15990910        Asp_cesB          10 : DPVCGMTVTEESKYH-EEFKGKTYYFCSDKCQSKFHSSP : 47
20090191        Mac_MA1330        14 : DPICGMPVDTEKAQFKAEIRGGTYYFCNEEHKRSFLENP : 52
22405416        Fac_Faci_p_309     4 : DPVCGMYV-SEDSKIYSDRDGTRYYFCSQGCKDKFDKPD : 41
13474443        Mlo_mlr5325       35 : DPVCGMTVDPAAGKPTSEHGGRLYHFCSERCRSKFQAEP : 73
13474443        Mlo_mlr5325       81 : DPVCGMSVDRATARHLVRHEGQGFYFCSAGCKAKFEAAP : 119
16082160        Tac_Ta1143         4 : DPVCGMYV-PETSDLYVDKDGQRYYFCSKGCMEKFLSPE : 41
22980562        Rme_Reut_p_5280   50 : DPVCGMAV-STESKFRAEHDGKQYYFCSNSCHQKFLQEP : 87
7531048         Rle_CTA_Rle       36 : DPICGMTVDPQAGKPSLGHGGRIYHFCSEHCRTKFAAAP : 74
17158728        Nsp_all7592       24 : DPICGMTVP-KATSLKTERGGRNYYFCSQTCLNTFL-DP : 59
7531048         Rle_ACTP          82 : DPVCGMSVDRSTARYFLKAEGEKFYFCSAACQAKFEADP : 120
13633955        Sty_SilP          45 : ESVCGMVILPDKAHSSIRYQDHQLYFCSASCESKFKAHP : 83
18482403        Sma_SilP          43 : DPVCGMAILPDRAHSSIRYQDHQLYFCSASCESKFKAHP : 81
22963632        Rpa_Rpal_p_2934   68 : DPVCGMTVDTATAQHRLDHDGQTYYFCCSGCRDTFSADP : 106
22963632        Rpa_Rpal_p_2934  161 : DPVCGMTVDVATSKHSFEHDGTTYHFCCGGCRTKFAADP : 199
22963632        Rpa_Rpal_p_2934  234 : DPVCGMKVDPATSKHRFAYKGTTYHFCREACQTKFAADP : 272
23109139        Nar_Saro_p_2361   39 : DPVCGMSVDPATTPHVATHDGAHHYFCSAGCLAKFKTDP : 77
23335646        Blo_Blon_p_679   858 : DPVCGMTVAVNADAITREYEGKSYYFCGEHCATNFMKAP : 896
22988795        Bfu_Bcep_p_6738  394 : NPVCGMAVDPASAKHVIDYGGERVYFCCDGCKLEFERAP : 432
22989416        Bfu_Bcep_p_7363  338 : NPVCGMAVEIASAKHVLDYGGQRIYFCCDCCKLEFERRP : 376
23053339        Gme_Gmet_p_261   182 : DLVCGMPISATTAPCRIELHGRALYFCSEHCKDAYLKEK : 216
23052649        Mba_Meth_p_3932    4 : DPICKKIISDNTEYFS-DYGGKSYYFCSPECKQKFDALE : 41
21227327        Mma_MM1225         4 : DPICKKIISRDTEYFS-DYGGKNYYFCSSECKHKFDALE : 41
20093305        Mac_MA4520         4 : DPICKKIIPENSEYIS-DYGGKTYYFCSPECKQKFDVLE : 41
22405415        Fac_Faci_p_308   128 : CDYCDSII-SGKPHILDANH-NKLYFCCETCKSEYIQNH : 164
16082226        Tac_Ta1218       136 : CDYCGKII-VSDPIIVHSHN-RDYYVCCPNCEHDLKKRL : 172
13541208        Tvo_TVN0377      134 : CDYCGNQI-HGDPISVKWKN-RTYLVCCPNCEKDMLKRL : 170
15622809        Sto_ST1716       125 : CDYCGNEI-KGNPYLVKLGK-KVYYTCCKTCQTQLKKKL : 161
15899376        Sso_SSO2652      118 : CDYCGKEI-YDNPLTYKVGR-KTYYACCNSCLSGLKEKF : 154
18977111        Pfu_PF0739       204 : CDYCGKEI-VGEPIVYKYHN-KVYFFCCPTCFREFKKAR : 240
11498085        Afu_AF0474       141 : CDYCGKEM-CDEPIVYRLKN-KVYVLCCKTCLREFKEIQ : 177
15790250        Hsp_VNG1179C     193 : CDECGNTVTSEGT--TATIDGDRHHFCCQSCERQFRQRY : 229
12644413        Hsa_DXS6673E     591 : CHYCHSLF-SGKPEVLDWQD-QVFQFCCRDCCEDFKRLR : 627
12644413        Hsa_DXS6673E     679 : CTYCSQTCQRGVT---EQLDGSTWDFCSEDCKSKYLLWY : 714
12644413        Hsa_DXS6673E     311 : CAHCRTPLQKGQT--AYQRKGLPQLFCSSSCLTTFSKKP : 347
12644413        Hsa_DXS6673E     498 : CVWCKTLCKNFEMLSHVDRNGKTSLFCSLCCTTSYKVKQ : 536
12644413        Hsa_DXS6673E     546 : CSFCRRSLSDPCY--YNKVDRTVYQFCSPSCWTKFQRTS : 582
12644413        Hsa_DXS6673E     406 : ATRCSICQKTGEVLHEVSNGSVVHRLCSDSCFSKFRANK : 444
12644413        Hsa_DXS6673E     720 : CHACKRQGKLLET---IHWRGQIRHFCNQQCLLRFYSQQ : 755
9790027         Mmu_ZFP261       593 : CHYCHSLF-SGKPEVLEWQD-QVFQFCCRDCCEDFKRLR : 629
6005978         Hsa_ZFP258       430 : CQHCNHLFATKPE--LLFYKGKMFLFCGKNCSDEYKKKN : 466
11292209        Hsa_FIM protein  641 : CNYCKNSFCSKPE--ILEWENKVHQFCSKTCSDDYKKLH : 677
13541282        Tvo_TVN0451        6 : CSFCGKTIEPGTGIMYVRKDGAILYFCSNKCKKNMIGLN : 44
15897177        Sso_rpl24E         7 : CSFCGHEIPPGTGLMYVRNDGTILWFCSSKCRKSMLKYH : 45
21592958        Ath_rpl30          6 : CWFCSSTIYPGHGIQFVRNDAKIFRFCRSKCHKNFKMKR : 44
20381079        Hsa_rpl30          6 : CYFCSGPIYPGHGMMFVRNDCKVFRFCKSKCHKNFKKKR : 44
132773          Hma_1FFK           6 : CDYCGTDIEPGTGTMFVHKDGATTHFCSSKCENNADLGR : 44
Secondary structure Hma_1FFK             E....EEE.....EEEEE...EEEEE..HHHHHHHH...
Consensus 80%                            sshCt..h.............p.bbbCs..C...b....
```

Figure 4.2 (previous page; NB: For COLOR version of this figure, see APPEDIX I) Multiple sequence alignment of TRASH domains that were identified using iterative PSI–BLAST searches and Hidden Markov profile searches[18]. The 80% consensus shown below the alignment was derived using the following amino-acid classes: hydrophobic [*h* (A,L,I,C,V,M,Y,F,W); *yellow* shading]; small [*s* (A,C,D,G,N,P,S,T,V); *green*] and the tiny subset of the small class [*t* (G,A,S); *green* shading]; polar [*p* (C,D,E,H,K,N,Q,R,S,T); *blue* text] and big [*b* (E,F,I,L,M,Q,R,W,Y); *gray* shading]. Completely conserved cysteine residues within the hydrophobic class are shown in red in the consensus. The secondary structure of TRASH was predicted using PHD[15] and is depicted above the alignment; it is in agreement with the known structure of the *Haloarcula marismortui* RPL24E[16], which is depicted under the alignment. β-strands and α-helices are represented by *e* and *h*, respectively; uppercase indicates the prediction has an accuracy >82%, lowercase represents an accuracy >72%. The limits of the domains are indicated by the position numbers on each side and the sequences are denoted by their GenBank Identifier, followed by the species abbreviation and gene name. Stand-alone versions of the TRASH domain that were missed in the initial genome annotation are indicated with an asterisk (protein sequences are available at http://www.ftns.wau.nl/micr/bacgen/TRASH/stand-alone.htm). This multiple sequence alignment (alignment number ALIGN_000512) has been deposited with the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000512).

centered around the invariant cysteine residues. According to the distant, but significant, sequence similarity with RPL24E, the TRASH domain is anticipated to adopt a 'treble-clef' fold[16,17] consisting of three or four β-strand followed by a C-terminal α-helix (Figure 4.2). This is in agreement with the secondary structure that was predicted by PHD. It is anticipated that metal coordination by TRASH is conducted in a similar manner to that proposed for Zn^{2+} coordination in RPL24E[17].

Domain architectures and functional roles of TRASH-containing proteins

The presence of TRASH in the N-terminal region of several P-type CTAs strongly suggests a functional role that is similar to that of the well-characterized heavy-metal associated (HMA) domain (PFAM00403). The HMA domain contains a 30-residue motif with two conserved cysteine residues that are involved in metal coordination; this motif has been found in several of these P-type CTAs and other metal detoxification proteins[19]. It was noted that in several phylogenetically distinct genomes, genes encoding TRASH-containing P-type CTAs are clustered with genes encoding stand-alone TRASH domains (Figure 4.1a). This suggests a functional link between these proteins that might be analogous to that of the HMA-containing CTAs and the HMA copper chaperones. The stand-alone TRASH domains that cluster with the P-type CTAs probably function as copper chaperones and are involved in metal-trafficking towards the target N-terminal TRASH domains of the CTAs, which subsequently translocate copper across the cell membrane. In addition, the metallotrafficking role of TRASH is suggested by the conserved clustering of (to date, undetected) genes encoding standalone TRASH proteins with genes encoding mercury reductase (MerA, SSO2689/ST1075) and the transcriptional regulator MerR (SSO2688/ST1076), which are part of the mercury detoxification regulon in *Sulfolobus solfataricus* and *Sulfolobus tokodaii*. It is anticipated that the stand-alone TRASH protein is involved in scavenging Hg^{2+} and subsequent trafficking towards the mercury reductase, which reduces Hg^{2+} to Hg^0. TRASH is also found in association with homologs of the β-subunit of cofactor F420-dependent hydrogenases in *Methanosarcina* (Figure 4.1b). In this instance, a role for TRASH is not obvious. The presence of the TRASH in the C-terminal region of Archaeal transcriptional regulators might suggest that TRASH functions as a metal-sensing regulatory module.

The vertebrate DXS6672E-like proteins that contain multiple TRASH domains have been reported to be located in the cell nucleus, and are subjected to differential splicing. Another member of this family, the human FIM protein (ZNF198/RAMP), is involved in atypical myeloproliferative disorder when its gene is fused to the gene encoding fibroblast growth factor receptor 1 (FGFR1)[20]. These findings show that TRASH is a versatile metal binding domain that is abundant in prokaryotes, but also in some eukaryotic proteins in different domain architectures (Figure 4.1b).

Implications for genome annotation

The described search for stand-alone versions of the TRASH domain in completely sequenced genomes revealed that the small ORFs have sometimes been overlooked during the annotation process. We have deposited these on http://www.ftns.wau.nl/micr/bacgen/TRASH/stand-alone.htm.

REFERENCES

1 Degtyarenko, K. (2000) Bioinorganic motifs: towards functional classification of metalloproteins, *Bioinformatics* **16**, 851-64

2 Camakaris, J., Voskoboinik, I. and Mercer, J.F. (1999) Molecular mechanisms of copper homeostasis, *Biochem Biophys Res Commun* **261**, 225-32

3 Blindauer, C.A., Harrison, M.D., Robinson, A.K., Parkinson, J.A., Bowness, P.W., Sadler, P.J. and Robinson, N.J. (2002) Multiple bacteria encode metallothioneins and SmtA-like zinc fingers, *Mol Microbiol* **45**, 1421-32

4 Lutsenko, S. and Kaplan, J.H. (1996) P-type ATPases, *Trends Biochem Sci* **21**, 467

5 Rosenzweig, A.C. (2002) Metallochaperones: bind and deliver, *Chem Biol* **9**, 673-7

6 Harrison, M.D., Jones, C.E., Solioz, M. and Dameron, C.T. (2000) Intracellular copper routing: the role of copper chaperones, *Trends Biochem Sci* **25**, 29-32

7 Rosenzweig, A.C. (2001) Copper delivery by metallochaperone proteins, *Acc Chem Res* **34**, 119-28

8 Wernimont, A.K., Huffman, D.L., Lamb, A.L., O'Halloran, T.V. and Rosenzweig, A.C. (2000) Structural basis for copper transfer by the metallochaperone for the Menkes/Wilson disease proteins, *Nat Struct Biol* **7**, 766-71

9 Larin, D., Mekios, C., Das, K., Ross, B., Yang, A.S. and Gilliam, T.C. (1999) Characterization of the interaction between the Wilson and Menkes disease proteins and the cytoplasmic copper chaperone, HAH1p, *J Biol Chem* **274**, 28497-504

10 Walker, J.M., Tsivkovskii, R. and Lutsenko, S. (2002) Metallochaperone Atox1 transfers copper to the NH2-terminal domain of the Wilson's disease protein and regulates its catalytic activity, *J Biol Chem* **277**, 27953-9

11 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**, 3389-402

12 Letunic, I., Goodstadt, L., Dickens, N.J., Doerks, T., Schultz, J., Mott, R., Ciccarelli, F., Copley, R.R., Ponting, C.P. and Bork, P. (2002) Recent improvements to the SMART domain-based sequence annotation resource, *Nucleic Acids Res* **30**, 242-4

13 Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. and Sonnhammer, E.L. (2002) The Pfam protein families database, *Nucleic Acids Res* **30**, 276-80

14 Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* **22**, 4673-80

15 Rost, B., Schneider, R. and Sander, C. (1997) Protein fold recognition by prediction-based threading, *J Mol Biol* **270**, 471-80

16 Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution, *Science* **289**, 905-20

17 Grishin, N.V. (2001) Treble clef finger--a functionally diverse zinc-binding structural motif, *Nucleic Acids Res* **29**, 1703-14

18 Eddy, S.R. (1998) Profile hidden Markov models, *Bioinformatics* **14**, 755-63

19 Jordan, I.K., Natale, D.A., Koonin, E.V. and Galperin, M.Y. (2001) Independent evolution of heavy metal-associated domains in copper chaperones and copper-transporting atpases, *J Mol Evol* **53**, 622-33

20 Popovici, C., Adelaide, J., Ollendorff, V., Chaffanet, M., Guasch, G., Jacrot, M., Leroux, D., Birnbaum, D. and Pebusque, M.J. (1998) Fibroblast growth factor receptor 1 is fused to FIM in stem-cell myeloproliferative disorder with t(8;13), *Proc Natl Acad Sci U S A* **95**, 5712-7

Chapter5

A novel type of copper-responsive transcription regulator in archaea

Thijs J. G. Ettema
Arjen B. Brinkman
Packo P. Lamers
Noor G. Kornet
Willem M. de Vos
John van der Oost

Using a comparative genomics approach, a copper resistance gene cluster has been identified in multiple archaeal genomes. The *cop* cluster is predicted to encode a metallochaperone (CopM), a P-type copper-exporting ATPase (CopA), and a novel, archaea-specific transcriptional regulator (CopT) which might control the expression of the *cop* genes. Sequence analysis revealed that CopT has an N-terminal DNA-binding helix-turn-helix domain and a C-terminal TRASH domain; TRASH is a novel domain which has recently been proposed to be uniquely involved in metal-binding in sensors, transporters and trafficking proteins in prokaryotes. The present study describes the molecular characterization of the *cop* gene cluster in the thermoacidophilic crenarchaeon *Sulfolobus solfataricus*. The polycistronic *cop*MA transcript was found to accumulate in response to growth-inhibiting copper concentrations, whereas *cop*T transcript abundance appeared to be constitutive. DNA-binding assays revealed that CopT binds to the *cop*MA promoter at multiple sites, both upstream and downstream of the predicted TATA-BRE site. Copper was found to specifically modulate the affinity of DNA binding by CopT. Furthermore, co-occupancy experiments suggest that CopT may stimulate TFB/TBP complex recruitment, whereas it impairs RNA polymerase recruitment. This study describes the first copper-responsive operon in archaea, a new family of archaeal DNA-binding proteins, and support a prominent role of the original TRASH domain in archaeal copper response. A model is proposed for copper-responsive transcriptional regulation of the *cop*MA gene cluster.

Submitted for publication

INTRODUCTION

Despite the fact that the basal transcription machinery of archaea resembles that of eukaryotes, their regulatory mechanisms appear to be different. Although a few homologs of eukaryote-like regulators are encoded by archaeal genomes, many potential bacterial-type transcriptional regulators have been identified[1]. Apparently, transcription regulation in archaea is a mixed heritage which involves a eukaryotic transcription machinery that is modulated by bacterial-like regulators. Several studies have revealed that most of these archaeal transcriptional regulators act as repressors of transcription[2-7]. Remarkably, recent studies also have provided evidence of positive control by these regulators[8,9]. In addition to bacterial-like regulators, archaea appear to contain archaea-specific regulators[10,11]. Recently, we have identified an archaea-specific transcription regulator that has been proposed to play a role in heavy metal resistance based on conserved gene context analysis[12]. Despite the fact that several archaeal species are able to thrive in environments that contain extremely high metal concentrations[13], information about heavy metal resistance in these organisms is limited[6].

Metal ions play an essential role as trace elements in many biological systems; current estimates indicate that over half of all proteins are metalloproteins, containing metal ions either as a structural component or as a catalytic co-factor[14]. However, at higher concentrations heavy-metal ions form unspecific complexes in the cell, leading to toxic effects. Some heavy-metal cations, like Hg^{2+}, Cd^{2+} and Ag^{+}, form strong toxic complexes, which makes them unsuitable for physiological function. Even essential trace elements like Zn^{2+} or Ni^{2+} and especially Cu^{2+} are toxic at elevated concentrations[15]. Both eukaryotes and prokaryotes have evolved several mechanisms that prevent the cell from being intoxicated by these metals. These homeostasis mechanisms include (i) sequestering of metals by metallothioneins[16,17], (ii) the efflux of metals by P-type cation-transporting ATPases (CTAs)[18], and (iii) trafficking of metal ions by metallochaperones[19,20].

Metal homeostasis has been well studied in bacteria and eukaryotes. In human, several diseases have been linked to an impaired metal balance, like the Menkes and Wilsons copper storage diseases[21]. In bacteria, several regulatory mechanisms have been identified that ensure a tight regulation of the expression of genes encoding metal homeostasis components. These mechanisms include two-component regulatory systems and metalloregulators. In general, the latter class of proteins comprise a DNA binding domain and a metal sensing domain. Examples of bacterial families of metalloregulators are the SmtB/ArsR transcriptional repressors[22], DtxR/MntR family[23] and the MerR family[24]. These metalloregulatory proteins have been found to display distinct metal selectivity profiles, in order to generate a metal-specific transcriptional response[22,23,25].

In this study we have investigated a putative copper (*cop*) resistance gene cluster in the thermoacidophilic crenarchaeon *Sulfolobus solfataricus*. This *cop* gene cluster, which is encoded by several archaeal genomes, encodes a novel transcriptional regulator, designated CopT, which contains a recently described metal-sensing domain, the TRASH domain. The results of this study suggest a prominent role for the TRASH domain in copper resistance in archaea.

RESULTS AND DISCUSSION

Analysis of the archaeal *cop* cluster

In an attempt to identify novel transcriptional regulators in archaea, we have screened the available archaeal genomes for the presence of DNA binding domains. Local gene context analysis of the obtained potential transcriptional regulators, revealed the existence of a conserved gene cluster that comprises a potential *cop* operon that is encoded by multiple archaeal genomes (Figure 5.1). Besides the archaea-specific Transcription regulator (*cop*T), the gene cluster consists of genes encoding a putative Metallochaperone (*cop*M)[12] and a P-type cation-transporting ATPase (*cop*A). The conserved gene organization of the genes that comprise the gene cluster strongly suggests a functional link between CopT, CopM and CopA, and probably that expression of the *cop*MA genes is controlled by CopT. The organization of genes encoding transcriptional regulators with their respective target genes in a divergon structure is a common organization for genes encoding transcriptional regulators and their targets[26] (Figure 5.1).

The archaea-specific *cop*T gene is found in the two major archaeal divisions, the crenarchaea and the euryarchaea. Domain architecture analysis[27] reveals that CopT contains an N-terminal HTH motif (Figure 5.2A) that resembles the DNA binding motifs of prokaryotic transcriptional regulators, such as MarR, IclR, and Lrp[28]. The presence of these typical prokaryotic HTH motifs has resulted in the mis-annotation of CopT proteins, as being members of the Lrp family of transcriptional regulators. However, CopT proteins lack a C-terminal RAM domain[29], which is the characteristic ligand-binding domain of Lrp-like proteins[28]. Instead, the C-terminal part of

CopT is a module that has recently been classified as TRASH domain[12]. The TRASH domain is a putative metal-binding motif, containing a conserved cysteine signature that is located within a 30-amino acid region (Figure 5.2B). The TRASH domain appears to be specifically involved in heavy metal resistance and is present in more than 140 proteins, distributed among all three domains of life. Based on distant homology with ribosomal protein RPL24E of *Haloarcula marismortui*, the TRASH domain has been anticipated to adopt a 'treble clef' fold[30] (Figure 5.2C). Metal coordination by TRASH is anticipated to occur in a similar fashion to that proposed for Zn^{2+} coordination in RPL24E[30]. Whereas the N-terminal HTH motif and the C-terminal extended TRASH domain are well conserved between archaeal CopT proteins, the central hinge regions show a higher degree of variability (Figure 5.2B).

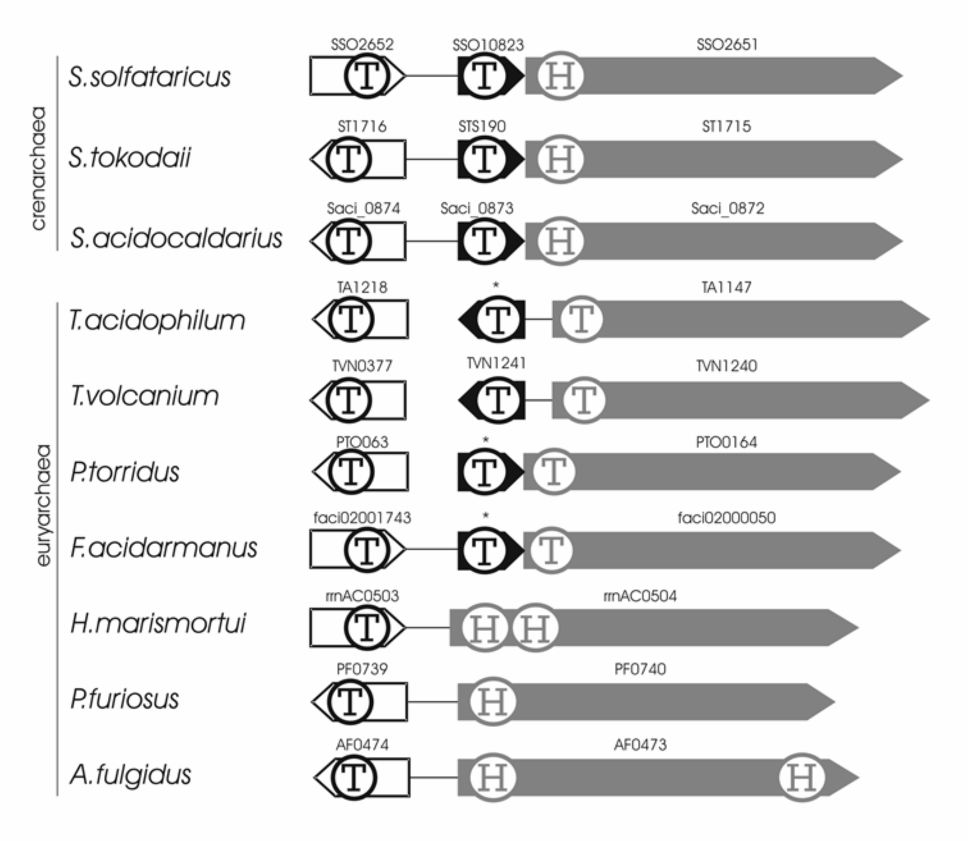
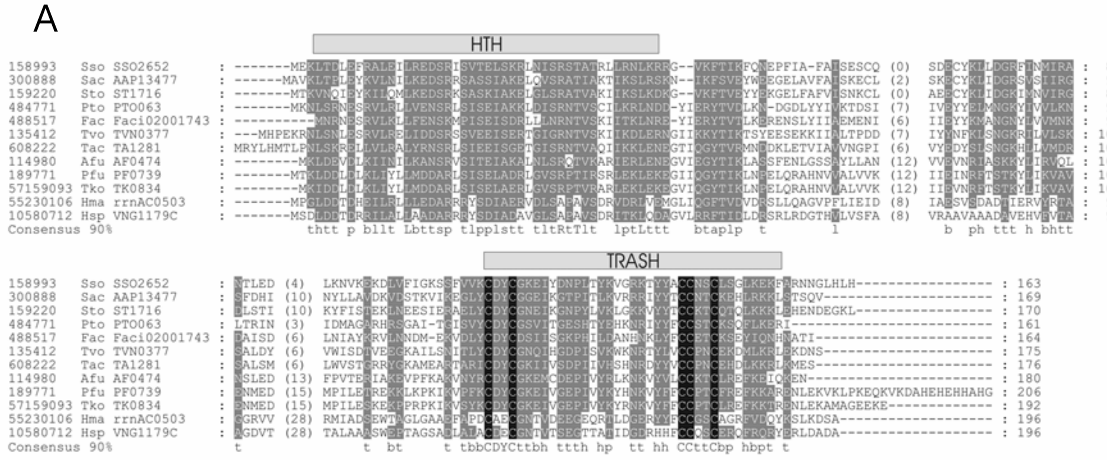


Figure 5.1 Conserved gene organization of archaeal *cop* regulons. Genes encoding CopT, CopM and CopA are indicated wit white, black and grey arrows respectively. Drawn lines indicate a physical linkage between genes. Gene numbers are indicated above the genes. Genes that have been overlooked in genome annotation and are absent from protein databases are indicated with an asterisk. Genes are approximately drawn to scale. Domain abbreviations: *T*: TRASH domain; *H*: HMA domain. (Adapted from reference [12]).

The archaeal CopA proteins that are encoded by the *cop* operons of different archaea are most similar to copper transporting ATPases. For example, CopA from *Sulfolobus solfataricus* (SSO2651) displays 20 to 32% identity with several well-characterized copper-exporting P-type ATPases of the bacteria *Enterococcus hirae*[31], *Escherichia coli*[32], *Streptococcus mutans*[33], and *Helicobacter pylori*[34], as well as with the human Menkes (MNK) and Wilsons (WD) disease-related proteins[35], suggesting that the archaeal CopA proteins are involved in copper transport. The well-studied CopA proteins, such as the human MNK and WD proteins, contain up to six N-terminal metal-binding domains that have been shown to be involved in metal transfer. These domains, called HMA domains (heavy metal associated), interact with metal-loaded metallochaperones. For example, the human metallochaperone HAH1[36] transfers copper to the N-terminal HMA domains of the Wilson's disease protein and regulates its catalytic activity[37,38]. These metallochaperones are homologous to the N-terminal metal binding domains of the metal transporting ATPases. Some of the archaeal CopA proteins encoded by the *cop* operon also contain N-terminal HMA domains (e.g CopA from *Sulfolobus* spp., *P.furiosus* and *A.fulgidus*) (Figure 5.1). Interestingly, the remaining archaeal CopA proteins lack these HMA domains, but instead contain N-terminal TRASH domains[12] (Figure 5.1). It should be noted that the occurrence of these TRASH-CopA proteins is not restricted to archaea. Several bacterial TRASH-CopA proteins have been identified, and these members include instances that contain multiple N-terminal TRASH domains[12]. In addition, the archaeal *cop* operon comprises a short gene (*cop*M) that encodes a stand-alone TRASH domain, which is anticipated to have a metallochaperone-like role[12].



B

TRASH domain

```
Secondary structure (PHD pred)      ......e........eeee...eEEEe.hhHHHHhh...
15899376    Sso SSO2652       118 : CDYCGKEI-YDNPLTYKVGR-KTYYACCNSCLSGLKEKF : 154
132773     Hma 1FFK             6 : CDYCGTDIEPGTGTMFVHKDGATTHFCSSKCENNADLGR : 44
Secondary structure (Hma 1FFK)      E....EEE.....EEEEE...EEEEE..HHHHHHHH...
```

HTH domain

```
Secondary structure (PHD pred)      ..HHHHHHHHHHH....HHHHHHHHHH....HHHHHHHHHHHH
158993  Sso SSO2652             4 : LTDLEFRALEILREDSRISVTELSKRLNISRSTATRLLRNLKR : 47
4416541 Pfu 1I1G                2 : IDERDKIILEILEKDARTPFTEIAKKLGISETAVRKRVKALEE : 45
Secondary structure (Pfu 1I1G)      ...HHHHHHHHHHHTTTTT.HHHHHHHHT...HHHHHHHHH..
```

C

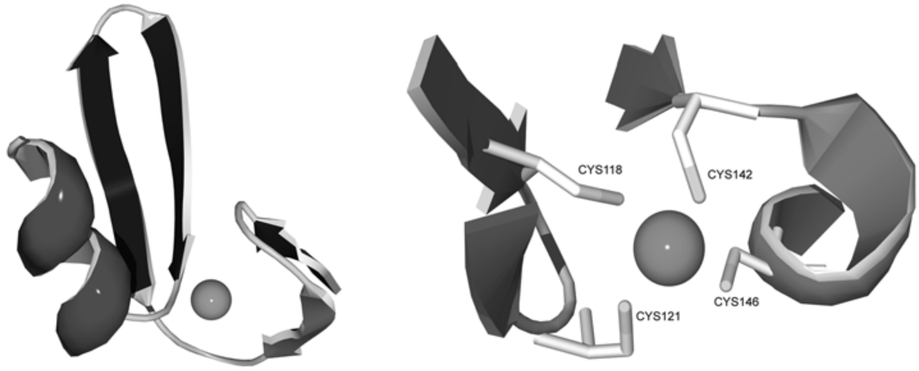


Figure 5.2 (NB: For COLOR version of this figure, see APPEDIX I) Sequence analysis of CopT. **(a, previous page)** Multiple alignment of the core of archaeal CopT proteins. The sequences are denoted as in figure 5.1 (except that for *S. acidocaldarius* CopT a locus entry is provided). The positions of the first and the last residue of the aligned region in the corresponding protein are indicated for each sequence. The 95% consensus sequence shown below the alignments was obtained using the following amino acid classes: aromatic (*a*: FYW), small (*s*: GASC), charged (*p*: STEDKRNQH), turn-forming (*t*: ASTDNVGPENRK), big (*b*: FILMVWYKREQ) and hydrophobic (*h*: ACFILMVWYH), of which the aliphatic subset (*l*: ILVA). Residues conserved > 95% are shown in *white* and boxed *gray*. Conserved cysteine residues that have been predicted to be involved in metal binding are shown in white and boxed black. The location of a putative helix-turn-helix motif and the TRASH domain are indicated above the alignment. The numbers within the alignment represent poorly conserved inserts that are not shown. For the CopT sequences of *P. furiosus* (PF0739) and *Halobacterium* sp. (VNG1179C), the transcription starts have been adjusted. **(b)** Structural alignment of the putative HTH domain of *S. solfataricus* CopT and the HTH domain of *P. furiosus* LrpA (*upper* panel), and of the TRASH domain of *S. solfataricus* CopT and the TRASH domain of *H. marismortui* RPL24E (*lower* panel). The sequences are denoted by Gene Identifier (gi) numbers from the GenBank database, species abbreviation and systematic gene numbers, respectively. Known structures of LrpA and RPL24E are depicted under the alignments. The predicted secondary structures of both domains of CopT were generated using PHD[39] and are depicted above the respective alignments. **(c)** Model of TRASH domain of *S. solfataricus* CopT based on homology modeling with the 3D structure of the ribosomal protein RPL24E of *H. marismortui*. *Left* panel: Overview of the modeled treble clef fold. *Right* panel: Close-up of the proposed metal binding site of the TRASH domain, based on the metal binding site of Hm-RPL24E. Cysteine residues that are anticipated to be involved in metal binding are numbered according to amino acid residue numbers of *S. solfataricus* CopT. Species abbreviations: Afu, *Archaeoglobus fulgidus*; Fac: *Ferroplasma acidarmanus*; Hma: *Haloarcula marismortui*; Hsp: *Halobacterium* sp. NRC1; Pto: *Picrophilus torridus*; Pfu: *Pyrococcus furiosus*; Sac: *Sulfolobus acidocaldarius*; Sso: *Sulfolobus solfataricus*; Sto: *Sulfolobus tokodaii*; Tko: *Thermococcus kodakaraensis*; Tac: *Thermoplasma acidophilum*; Tvo: *Thermoplasma volcanium*.

Expression of *cop* genes in *S. solfataricus*

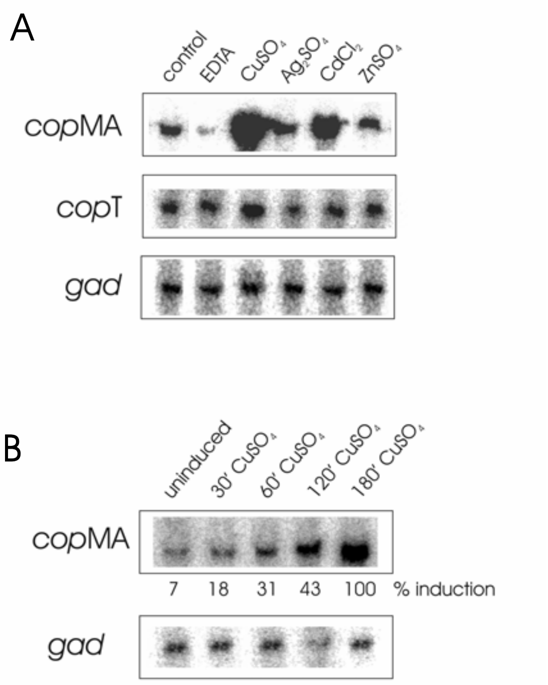


Figure 5.3 Effect of addition of different metals on expression levels of *cop*T and *cop*MA **(a)**. Minimal inhibiting concentrations of EDTA and various metals (see Table I) were added to an exponentially growing *S.solfataricus* culture. EDTA, 1 mM; $CuSO_4$, 5 mM; Ag_2SO_4, 8 μM; $CdCl_2$, 4 mM; $ZnSO_4$, 75 mM. Four hours after addition of EDTA or metals total RNA was isolated and analyzed using primer extension with *cop*T and *cop*MA antisense primers. **(b)** Induction of *cop*MA transcription by $CuSO_4$ was further studied at different time intervals (0, 30, 60, 120 and 180 min after addition of 5 mM $CuSO_4$). As an internal control, the *gad* (encoding *S.solfataricus* gluconate dehydratase) primer extension product is shown.

Table I. Minimal inhibitory concentrations of various metal salts for *S. solfataricus*

	MIC (μM)	
Compound	**Grogan**[38]	**This study**
EDTA	-	<1000
$CuSO_4$	5000	5000
Ag_2SO_4	8	8
$CdCl_2$	2000	>4000
$ZnSO_4$	50000	>75000
$NiSO_4$	600	600

The *cop* operon of *S. solfataricus* consists of the tandem orientated *cop*T gene and the *cop*MA gene pair (Figure 5.1). To determine the role of the genes of this cluster, the expression of the *cop* genes was analyzed under various conditions. First, the minimal inhibitory concentration (MIC) of different metals was determined for *S. solfataricus* (Table I). The obtained MIC values resembled the values reported by Grogan[40], except for Cd^{2+} and Zn^{2+}, which were found to be slightly higher (Table I).

In order to analyze the transcription regulation of the *cop* operon, different metals were added to a *S. solfataricus* culture that was growing exponentially in defined medium, at a concentration that was equivalent to the determined MIC value. Subsequently, total RNA was isolated and analyzed using primer extension analysis using antisense primers for *cop*T and *cop*MA. As shown in Figure 5.3A, transcription of *cop*T is constitutive under the tested conditions, whereas transcription of *cop*MA messenger is induced specifically in the presence of Cu^{2+} and Cd^{2+}. Under non-induced conditions, *cop*MA appears to be transcribed at a basal level, which decreases when the metal chelating agent EDTA was added. Contrarily, addition of EDTA has no apparent effect on *cop*T expression. It should be noted that addition of 1 mM EDTA immediately abolished growth of *S. solfataricus*. Induction of *cop*MA transcription by Cu^{2+} was further studied by isolation of total RNA from a *S. solfataricus* culture at different time intervals. Transcription of *cop*MA was readily induced after 30 min and increased during the 180 min after induction to a level more than 14-fold higher than the basal transcription level (Figure 5.3B). To determine whether basal *cop*MA transcription is due to the small amount of Cu^{2+} (0.45 μM) that is present in the defined medium, *cop*MA transcription was also determined in defined medium from which Cu^{2+} was omitted. Although no significant decrease in basal transcription was observed (not shown), we cannot rule out the possibility that trace amounts of Cu^{2+} were still present in the demineralized water that was used to prepare the defined medium. We conclude that the genes in the *cop* regulon in *S. solfataricus* are most likely

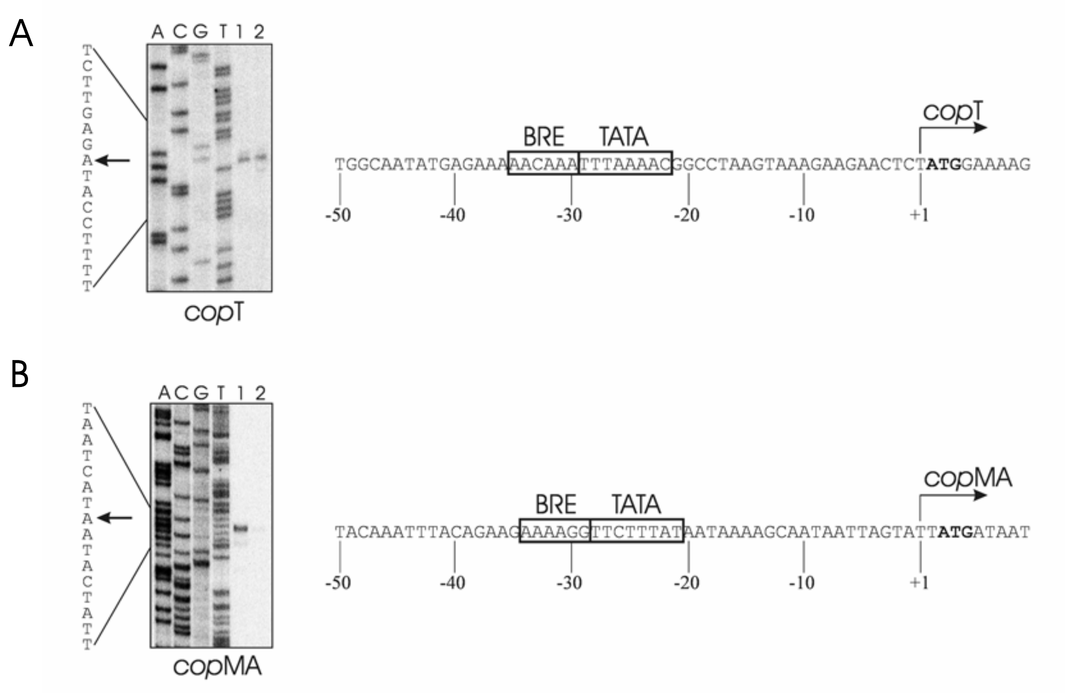


Figure 5.4 Mapping of the transcription start sites of *cop*T and *cop*MA genes. Primer extension products of *cop*T **(a)** and *cop*MA **(b)** were analyzed on a sequencing gel along a sequence ladder. Lane 1 and 2 contain primer extension products using RNA from $CuSO_4$-induced cells and non-induced cells respectively. The transcriptional start sites of *cop*T and *cop*MA are indicated with an arrow (+1), and presumptive BREs and TATA-boxes are indicated by rectangles.

involved in the efflux of copper and cadmium, since these metals specifically induce expression of the *cop*MA transcript. As with many transcriptional regulators, *cop*T expression appears to be maintained at a constitutive level under the tested conditions.

To determine the transcription start sites of the *cop*T and *cop*MA transcripts, we separated the obtained primer extension products on a larger gel along with sequence reactions that were performed with the same *cop*T and *cop*MA antisense primers. As shown in Figure 5.4A, *cop*T transcription is initiated one basepair upstream of the start codon of the *cop*T gene, whereas transcription of *cop*MA is initiated two basepairs upstream of the *cop*M start codon (Figure 5.4B). We could not detect bands corresponding to a smaller primer extension product (results not shown), suggesting that the *cop*M and *cop*A genes are transcribed as a polycistronic mRNA, initiated from a single promoter upstream of *cop*M. Sequence elements matching *S. solfataricus* consensus TATA-boxes and BREs[41] are present upstream of the two mapped transcriptional starts (Figure 5.4A and B).

CopT specifically targets the *cop*MA promoter

The *cop*T gene was cloned into pET24d, and functionally overproduced in *E. coli*. Subsequenly CopT was purified to electrophoretic homogeneity as described in the experimental procedures (Figure 5.5A). To prevent undesired protein aggregation, the purified CopT protein was stored under anoxic conditions in the presence of excess amounts of DTT (10 mM). Subsequently, purified CopT was used in electrophoretic mobility shift assays (EMSAs) to determine whether it binds to the mapped *cop*T and *cop*MA promoters (P_{copT} resp. P_{copMA}, Figure 5.5B). CopT was unable to bind to the 200 bp P_{copT} fragment under the tested conditions (Figure 5.5C). This finding is in agreement with the observation that *cop*T expression is constitutive under all tested conditions and thus rules out the possibility that CopT is involved in regulation of its own expression. In contrast, transcriptional regulators of well characterized *cop* regulons of *E. coli* (CueR)[42,43] and *E. hirae* (CopY)[44] are subjected to autoregulation.

Binding of CopT was observed to P_{copMA}, a 222-bp DNA fragment containing the *cop*MA promoter (Figure 5.5D and E). EMSA experiments revealed that after the first CopT-DNA complex (complex 1) was formed, 4 additional complexes of distinct electrophoretic mobility were formed with increasing CopT concentration (Figure 5.5D). This suggests that multiple CopT binding sites are present on P_{copMA}. Further analysis of the different

CopT-DNA complexes revealed sigmoidal binding curves, which indicates that binding of CopT to these sites is of cooperative nature (Figure 5.5E)[45].

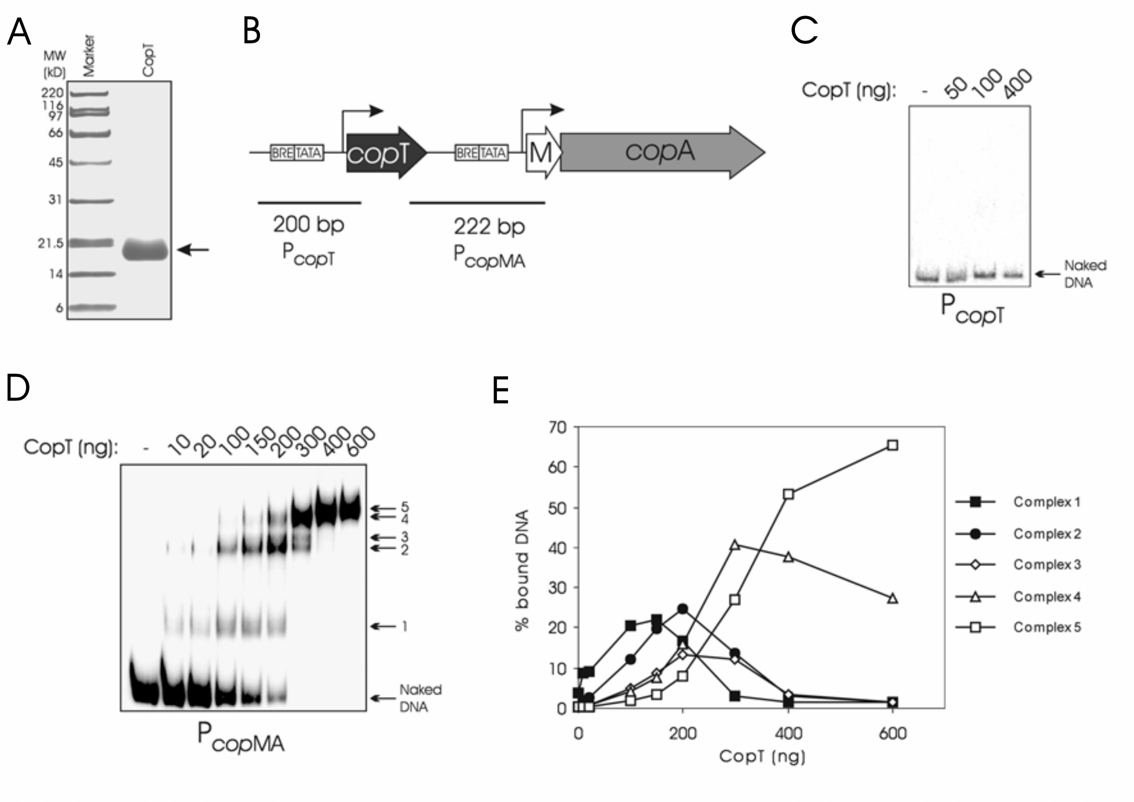


Figure 5.5 CopT binds to the *cop*MA promomoter, but not to its own promoter.**(a)** SDS-PAGE analysis of purified CopT. The *arrow* indicates a single band that corresponds with the size of a CopT monomer. **(b)** Schematic display of the DNA fragments used for electrophoretic mobility shift assays (EMSA), P_{copT} and P_{copMA}. Genes are not drawn to scale. **(c)** EMSA with P_{copT} using increasing amounts of CopT. **(d)** EMSA with P_{copMA} using increasing amounts of CopT. The five different CopT-DNA complexes are indicated with *1, 2, 3, 4* and *5* respectively. **(e)** Graph showing cooperative binding of CopT to P_{copMA}. Data is obtained by quantification of the band intensity of different CopT-DNA complexes as shown in Figure 5.5D.

Cu^{2+} modulates the CopT-DNA interaction

Because transcription analysis showed that both Cu^{2+} and Cd^{2+} induce expression of the *cop*MA operon, we investigated the effect of adding metal ions on formation of CopT-DNA complexes. Addition of Cu^{2+} to the binding reaction was found to reduce DNA binding capacity of CopT (Figure 5.6A), whereas addition of copper does not affect the binding of an unrelated archaeal transcriptional regulator (LrpA[5] to its target promoter P_{LrpA}, see Figure 5.6B); this strongly suggests a specific effect of copper on CopT functionality. Although Cd^{2+} induced *copMA* expression, addition of Cd^{2+} only has marginal effects on CopT-DNA complex formation, as reflected by a slight decrease of complex 1 (Figure 5.6A).

To study the effect of Cu^{2+} on CopT-DNA complex formation in more detail, EMSA analysis was performed with P_{copMA} and increasing amounts of CopT either in the presence or absence of Cu^{2+} (Figure 5.6C). Addition of Cu^{2+} to the binding reaction affected CopT-DNA binding in two ways. First, the overall binding affinity decreased, and second, the formation of complex 1 (Figure 5.6C) was completely obliterated. It is proposed that exogenous Cu^{2+} binds to the TRASH domain, to cause an allosteric conformational change in CopT, resulting in an altered DNA-binding mode involving lower DNA-binding affinity and perhaps different multimerisation properties. Consequently, this enables initiation of transcription of the *cop*MA genes. Together with the observed induction of *cop*MA transcription upon exposure to excess levels of exogenous copper *in vivo*, it is suggested that CopT is a repressor of *cop*MA transcription and that transcription occurs by Cu-CopT mediated derepression.

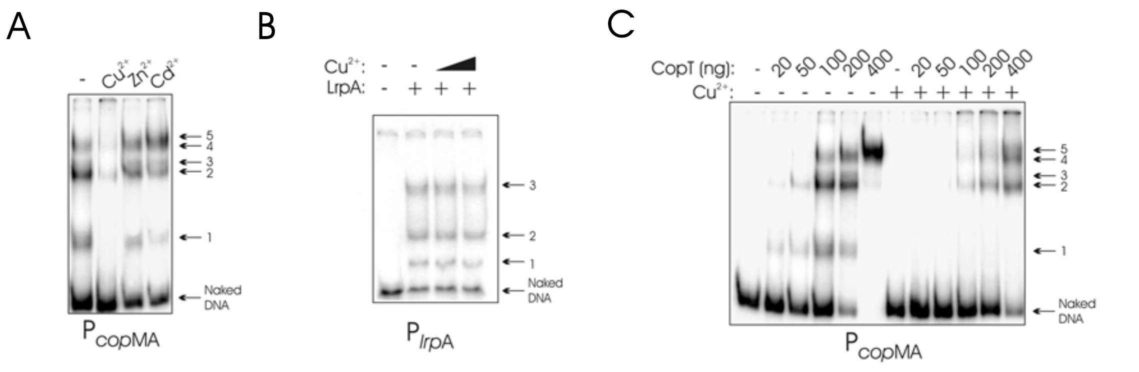


Figure 5.6 CopT-DNA complex formation is effected by $CuSO_4$. **(a)** Effects of various metal-ions on CopT-DNA binding. Binding reactions were performed as described in the methods and contained 200 ng of CopT. Metal salts were added to a final concentration of 100 μM. **(b)** ontrol experiment showing that addition of different concentrations of copper ions (0, 10 and 100 μM of $CuSO_4$) to an unrelated archaeal transcriptional regulator (LrpA) has no effect on binding to its target promoter P_{lrpA}. **(c)** Effect of Cu^{2+} on CopT-DNA binding at different concentrations of CopT (as indicated). Assays were performed in the presence or absence of 100 μM $CuSO_4$.

CopT binds to multiple binding sites at the *cop*MA promoter

To determine the locations where CopT interacts with the *cop*MA promoter, a DNase I footprint has been performed (Figure 5.7A and B). CopT protects P_{copMA} at multiple locations, suggesting the presence of multiple CopT binding sites. However, close examination of the protected regions did not reveal the presence of any type of CopT consensus motif. In addition, no consensus motif has been detected in the promoter regions of *cop* regulons in other archaea (data not shown). The absence of apparent sequence motifs either implies the existence of degenerative *cop* motif, or that specific DNA recognition by CopT is accomplished by other determinants, such as DNA conformation.

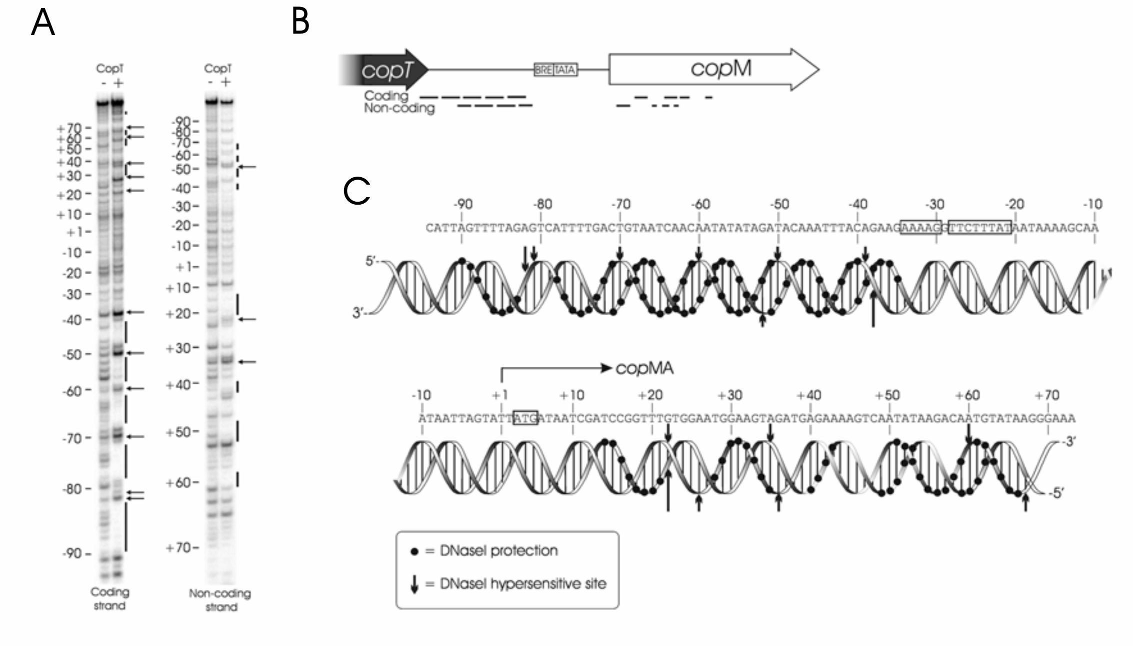


Figure 5.7 DNAseI footprint analysis of CopT bound to P_{copMA}. **(a)** DnaseI footprint assay on coding (*left panel*) and non-coding strand (*right panel*) using either a saturating amount (*+, right lane*) or no CopT (-, *left lane*) in the binding reaction as described in the methods. Sequence positions relative to the transcription start site (+1) are indicated *left* of the footprint. Protected areas are indicated by *vertical bars*. DNaseI hypersensitive sites are indicated by a *horizontal arrow*. **(b)** Schematic display of the CopT induced protected regions of P_{copMA}. The TATA-box and BRE element are shown. **(c)** Imposition of the DNaseI footprinting results on a double helical representation of the *cop*MA promoter region. *Filled circles* indicate sites that were protected during DNaseI degradation and *bold arrows* indicate DNaseI hypersensitive sites. The sequence of the non-template strand is indicated above the helical representation and boxed sequences indicate the TATA-box, BRE element and start codon (ATG) respectively. The sequence positions are relative to the *cop*MA transcription start site (+1), which is indicated by an *arrow*.

The regions that were protected by CopT in the DNase I footprint assay are located both upstream and downstream of the TATA box and BRE sequence, but are not overlapping with these elements. This suggests that potentially, the recruitment of basal transcription factors TFB and TBP is not prevented by the presence of CopT on P_{copMA}. To test this hypothesis we performed EMSAs using CopT, TBP-TFB and RNA polymerase. Interestingly, binding of TBP-TFB was only observed in the presence of CopT. (Figure 5.8), suggesting that CopT-DNA binding is a prerequisite for TBP-TFB recruitment. However, such CopT-TBP-TFB-DNA complex seems to impair recruitment of the RNA polymerase. This is in agreement with the results that were observed in the DNA footprinting assay, which shows binding of CopT in regions downstream of the transcription start point (Figure 5.7B and C). Since Cu^{2+} induces an altered binding mode of CopT to DNA, it is possible that in the presence of Cu^{2+} RNAP is recruited to the preformed CopT-TBP-TFB-DNA complex. Unfortunately, experiments to study such Cu^{2+}-induced RNA polymerase recruitment were hampered by Cu^{2+}-induced precipitation of TFB (not shown). Possibly, the added copper ions result in the displacement of zinc ions from the N-terminal zinc ribbon domains of TFB, causing the protein to precipitate. A similar finding has been reported by Blum and co-workers, who described TFB inactivation upon exposure to mercuric ions[46].

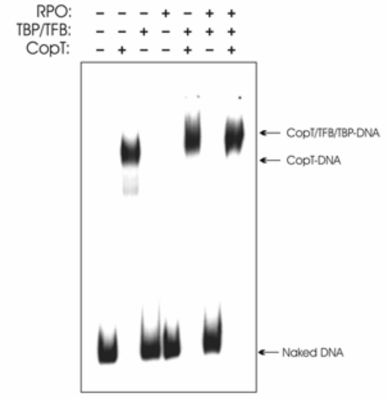


Figure 5.8 CopT induces TFB/TBP recruitment but prevents RNA polymerase recruitment.EMSA performed on the *cop*MA promoter using 200 ng of TBP and TFB, 250 ng of RNA polymerase (RPO) and/or 300 ng CopT, as indicated. The positions of the naked DNA and CopT–DNA and CopT/TBP/TFB–DNA complexes are shown.

A previous study in which the mode of action of an archaeal metalloregulator could be resolved in detail concerns the analysis of *A. fulgidus* MDR1[3]. This transcriptional repressor has been shown to recognize three operator sites on a promoter of a metal ABC transporter gene cluster in the presence of various metals *in vitro* and *in vivo*. Also here, repression was found to occur by impairing recruitment of RNA polymerase, while TFB and TBP were allowed to bind to their target sequences, governing a rapid transcriptional response after metal exposure[41].

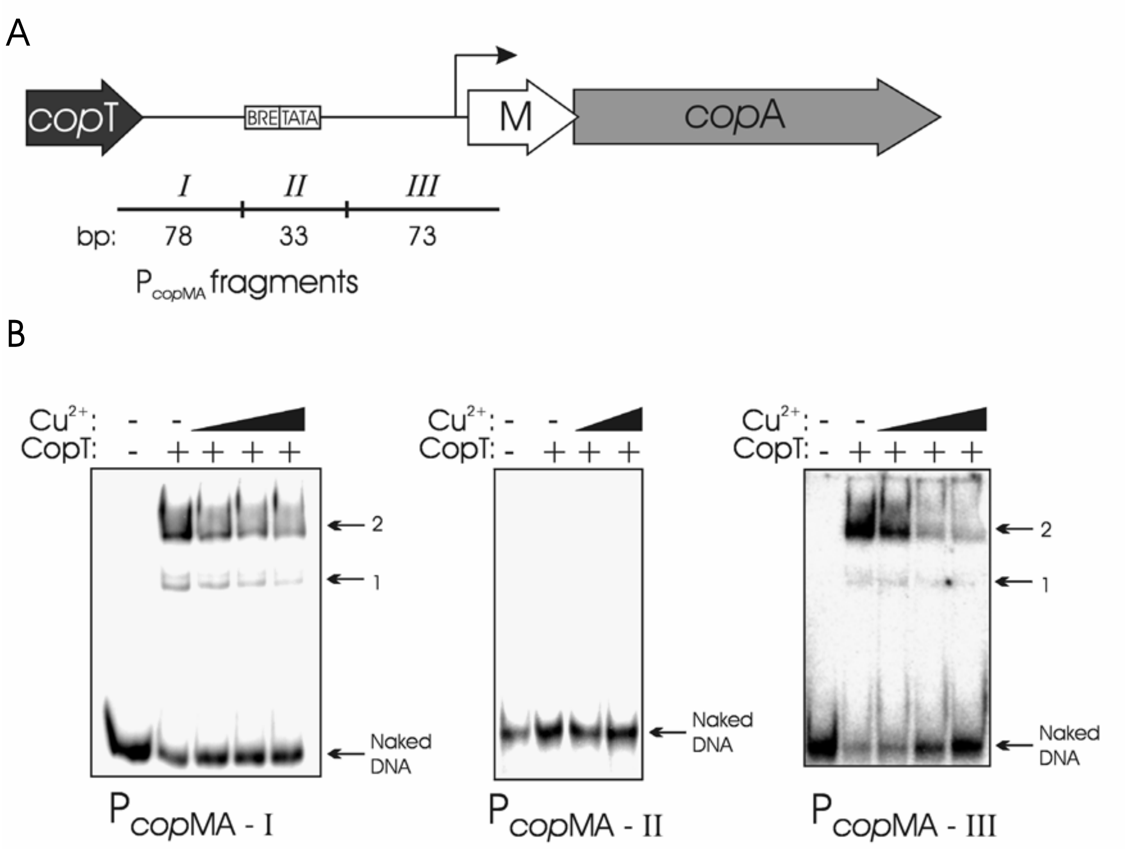


Figure 5.9 Binding of CopT to *cop*MA promoter fragments up and downstream of the TATA-box and BRE element is differentially influenced by exogenous Cu^{2+}. **(a)** Schematic display of the different DNA fragments used for EMSA analysis (P_{copMA}–I, II and III). **(b)** EMSA with CopT, P_{copMA} fragments and $CuSO_4$. If present, 400 ng CopT was added to the binding reactions. The different CopT-DNA complexes are indicated. $CuSO_4$, if present, was added to a final concentration of 10, 100 or 500 µM. For fragment P_{copMA}-II, the assay containing 100 µM $CuSO_4$ was not performed.

In order to study differential binding affinity of CopT for binding sites up and downstream of the TATA-BRE region, EMSAs were performed with different sub-fragments of P_{copMA} (Figure 5.9A). Indeed, CopT only bound to the fragments upstream and downstream of this TATA-BRE region but not to a 33-bp P_{copMA} fragment that includes the TATA-BRE site (Figure 5.9B fragment II). This is in agreement with the results from the DNase I footprint assay. Addition of saturating amounts of CopT resulted in two complexes for each promoter fragment (Figure 5.9B, fragment I and III). Addition of exogenous copper to the binding reaction appears to have a more severe effect on CopT binding to fragment III than to fragment I, where only a marginal effect can be observed. Apparently, CopT displays differential binding affinity to the sites on fragment I and III. It is tempting to speculate that the CopT binding site(s) on fragment III are mainly involved in transcriptional repression, which is released upon response to copper. Because CopT binding to the region upstream of the TATA-BRE fragment (fragment I) is less affected by exposure to copper, binding of CopT in this region may be more constitutive *in vivo*, where it could enforce a stimulatory affect on transcription, possibly by interacting with basal transcription factors. This would be in good agreement with the observations in Figure 5.8, which shows that the TFB-TBP complex only associates with P_{copMA} in the presence of CopT.

The TRASH domain is a central player in archaeal copper resistance

The results of this study hint at a central role of the TRASH domain in archaeal copper resistance. The conserved gene organization of TRASH-encoding *cop* genes in archaeal genomes implies a functional connection between the encoded proteins (Figure 5.1). Here, we show that the *cop*MA genes of *S. solfataricus* are specifically induced upon exposure to excess levels of exogenous copper and cadmium. Furthermore, CopT binds to the *cop*MA promoter *in vitro*, and this interaction is modulated upon addition of copper. Most likely, coordinated binding of copper ions by the C-terminal TRASH domain of CopT will result in a change in structural conformation, resulting in an altered DNA binding mode which involves reduced DNA-binding affinity and perhaps different a multimeric state. Thus, our results indicate that CopT is a transcriptional repressor of the *cop*MA operon and transcription is derepressed in the presence of exogenous copper ions. Increased levels of *cop*MA expression will result in increased levels of CopA and CopM protein. Whereas CopA will most probably be involved in copper efflux[47], the role of CopM remains unclear. In analogy with HMA-type metallochaperones, they might be employed with (i) metallo-trafficking-like functions[36,48,49], (ii) regulation of CopA activity[37,38] or, (iii) they could be involved in modulation of transcription regulation[50,51,52].

As can be observed from the diversity of metal binding domains of TRASH and HMA domain containing proteins in archaea and bacteria, these domains constitute evolutionarily mobile metal binding domains that are functionally, but not structurally, equivalent. It will be a major challenge to further study the roles of the individual proteins encoded by the archaeal *cop* clusters, as well as their interactions, in order to elucidate the mechanism that governs TRASH-domain mediated copper resistance in archaea.

EXPERIMENTAL PROCEDURES

Multiple alignment, domain analysis, secondary structure prediction and 3D modeling of *S. solfataricus* CopT

All protein sequences were retrieved from the Entrez database (NCBI). A multiple alignment of archaeal CopT protein sequences was constructed using the ClustalX program[53], followed by manual adjustment for conserved motifs based on BLAST results. Domain analysis of *S. solfataricus* CopT was performed using SMART[54] and protein secondary structure was predicted using the PHD program[39] as described previously[12]. Homology modeling of the TRASH domain of SSO2652 was performed by the ExPaSy homology-modeling server SWISS-MODEL (http://www.expasy.org/swissmod/SWISS-MODEL.html)[55] using the structure of *Haloarcula marismortui* RPL24E[56] as a template. Protein databank (PDB) files were visualized using SWISS-PDB viewer version 3.7, service pack 5[57] and rendered using POV-Ray version 3.6 (http://www.povray.org/).

Growth of *S. solfataricus*

S. solfataricus P2 (DSM1617) was grown in defined medium containing 3.1 g/l KH_2PO_4, 2.5 g/l $(NH_4)_2SO_4$, 0.2 g/l $MgSO_4{\cdot}7H_2O$, 0.25 g/l $CaCl_2{\cdot}2H_2O$, 1.8 mg/l $MnCl_2{\cdot}4H_2O$, 4.5 mg/l $Na_2B_4O_7{\cdot}10H_2O$, 0.22 mg/l $ZnSO_4{\cdot}7H_2O$, 0.06 mg/l $CuCl_2$, 0.03 mg/l $Na_2MoO_4{\cdot}2H_2O$, 0.03 mg/l $VOSO_4{\cdot}2H_2O$, and 0.01 mg/l $CoCl_2$, supplemented with 0.02 g/l $FeCl_3$, and vitamins, adjusted to pH 3.0 with H_2SO_4. After addition of sucrose as carbon source to a final concentration of 0.4% (w/v) and subsequent inoculation, the culture was propagated at 80°C, shaking at 130 rpm. For determination of the MIC values of the different metals, various concentrations of the metal salts were added to exponentially growing *S. solfataricus* cells as indicated at OD_{600} of 0.3-0.4, and growth was monitored by measuring the OD_{600}.

RNA isolation from *S. solfataricus* and primer extension analysis

S. solfataricus total RNA was isolated from mid-log cultures (A_{600} of ~ 0.5) grown as indicated using the RNeasy kit (QIAgen). 50 ml of culture was washed in 1 ml of medium and resuspended in 100 μl of TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0). After addition of 5 μl of 10% Triton X-100, the RNA was further purified according the manufacturer's prescriptions, except that genomic DNA was sheared through a 0.45-mm needle before the sample was applied onto a spin column. Columns were eluted twice with 50 μl of water. For the determination of the transcription start sites of *cop*T and the *cop*MA messenger, primer extension analysis was performed using the following radiolabeled antisense oligonucleotides: BG1131 (5'-GTGCTCCTACTG-ATATTAAGCC-3') for *cop*T, and BG1130 (5'-CATGTTGCACAATGCATCCC-3') for *cop*MA. For determination of *cop*T and *cop*MA expression levels upon addition of various metal salts (Figure 5.3A and B), the same oligonucleotides were used. As an internal control for RNA levels, an anti-sense oligonucleotide for the *gad* gene (BG2046, 5'-CAGATATAACTCTTAGTGTGGGTAC-3') was used[58], for which expression is expected to be unaffected upon addition of metal salts.

For the primer extension reaction, 10 μg of total RNA and 2.5 ng of radiolabeled oligonucleotide were resuspended in 2× AMV-RT buffer (Promega) in a final volume of 25 μl. Samples were heated to 70°C for 10 min and slowly cooled to room temperature. $MgCl_2$, dNTPs, RNasin, and AMV-RT (Promega) were added to a concentration of 5 mM, 0.4 mM, 0.8 units/μl, and 0.4 units/μl, respectively, in a final volume of 50 μl. The samples were incubated at 42°C for 30 min, extracted with phenol/chloroform, precipitated with ethanol and resuspended in formamide loading buffer. The primer extension product was analyzed on an 8% denaturing sequencing gel along with a sequence ladder that was generated using the same radiolabeled oligonucleotides. Primer extension products were quantified using Imagequant software (BioRad).

Recombinant production and purification of CopT

The *cop*T gene was PCR-amplified from *S. solfataricus* genomic DNA using the primers BG864 (5'-CGCGCCATGGAAAAGTTGACAGATTTA GAGTTTAG-3') and BG865 (5'- CGCGCGGATCCTAATGTAAGTGCAAGCC ATTGTTG -3'), which contain *Nco*I and *Bam*HI restriction sites, respectively (underlined). The generated PCR fragment was cloned into *Nco*I- *Bam*HI digested pET24d expression vector (Novagen) resulting in pWUR59. The sequence of the cloned *cop*T gene was verified by dideoxy sequencing and was subsequently transformed to the *E. coli* expression strain JM109(DE3)-pRIL (Novagen) to produce CopT protein. *E. coli* cells harboring pWUR59 were grown at 37°C in one liter of LB medium to an OD_{600} of 0.5, and CopT expression was induced by the addition of IPTG to a final concentration of 0.4 mM. After an overnight incubation at 37°C the cells were harvested by centrifugation for 10 min. at 5000 X g. Routinely, a bacterial cell pellet derived from 500 ml of expression culture was resuspended in 20 ml buffer P (20 mM Tris pH 7.0, 10 mM DTT) containing one Complete Mini protease inhibitor cocktail tablet (Roche) and subjected to cell lysis by sonification. After cell lysis, the cleared cell lysate was subjected to a heat treatment (30 min. at 80°C) and subsequently centrifuged for 30 min. at 16,000 X g. Recombinant CopT was then purified to apparent homogeinity by size exclusion chromatography using a Superdex 200 10/300 GL column (Amersham Biosciences), which was pre-equilibrated with buffer S (20 mM Tris pH 7.0, 0.1 M NaCl, 10 mM DTT). Fractions containing pure CopT were pooled and stored under anoxic conditions at 4°C. Typically, a 500-ml expression culture harvested approximately 5 mg of electrophoretically pure CopT.

Recombinant LrpA from *P. furiosus* was produced and purified as described before[5].

Electrophoretic Mobolity Shift Assays and DNAseI footprinting

DNA probes for EMSA were generated using PCR with the primer pairs BG1081 (5'-ACTAGTTGGATGGATATTAGGAATAGC-3') and BG1082 (5'-TCTCTTAAAATCTCCAGCGCTC-3') for $\mathrm{P}_{cop\mathrm{T}}$, primer pairs BG1079 (5'-TGCACGCAACAATGGCTTGC-3') and BG1725 (5'-GACAATGAGATGAGCAGAAATAG-3') for $\mathrm{P}_{cop\mathrm{MA}}$ and primer pairs BG1770 (5'-CAATGGCTTGCACTTACATTA-3'), BG1771 (5'-CTTCTGTAAATTTGTATCTATATA-3') for $\mathrm{P}_{cop\mathrm{MA}}$-I, primer pairs BG1772 (5'-GAAAAGGTTCTTTATAATAAAAG-3') and BG1773 (5'-CTAATTATTGCTTTTATTATAAAG-3') for $\mathrm{P}_{cop\mathrm{MA}}$-II, and primer pairs BG1774 (5'-GATAATCGATCCGGTTTGTGG-3') and BG1775 (5'-GATTTTTCCCTTATACATTGTC-3') for $\mathrm{P}_{cop\mathrm{MA}}$-III, respectively. PCR products were end-labeled using T4 kinase and radioactive $[\gamma\text{-}^{32}\mathrm{P}]$ATP (Amersham Biosciences) and purified from a 6% polyacrylamide gel. Binding reactions were performed in binding buffer B (50 mM Tris/pH 8.0, 1 mM DTT, 5% glycerol, and 5 ng/μl poly(dI.dC).poly(dI.dC)) and reactions were incubated at room temperature ($\mathrm{P}_{cop\mathrm{MA}}$-I - P $\mathrm{P}_{cop\mathrm{MA}}$-III) or at 50 °C (for $\mathrm{P}_{cop\mathrm{T}}$ and $\mathrm{P}_{cop\mathrm{MA}}$) for 20 min. Metals, if present, were added to a final concentration as indicated. Obtained protein-DNA complexes were separated on a non-denaturing 12% ($\mathrm{P}_{cop\mathrm{MA}}$-I - $\mathrm{P}_{cop\mathrm{MA}}$-III) or 6% (for $\mathrm{P}_{cop\mathrm{T}}$ and $\mathrm{P}_{cop\mathrm{MA}}$) polyacrylamide gel, buffered in 1× TBE buffer. Gels were dried, exposed to phosphor screens, and analyzed. EMSA using *P.furiosus* LrpA were performed as described before[5].

The promoter co-occupancy experiment shown in Figure 5.8 was performed as described by Bell *et al*[3], with the following modifications. The EMSA reactions were performed in Buffer B and contained 200 ng of TBP and

TFB, 250 ng RNAP (a kind gift from Dr. Steve Bell), and/or 300 ng CopT, as indicated. The reactions were electrophoresed on a 4.5% non-denaturing polyacrylamide gel, as described above.

For DNase I footprinting a P_{copMA} probe was generated using PCR with the oligonucleotides BG1079 and BG1725, by end-labeling one of the two oligonucleotides using T4 kinase and radioactive [γ-^{32}P]ATP. Probes were purified from a 6% polyacrylamide gel. The binding reactions contained 1 μg CopT and were performed at 50 °C in a total volume of 50 μl containing 50 mM Tris/pH 8.0, 25 mM MgCl2, 75 mM KCl, 1 mM DTT. After 20 min, the reaction was cooled to 48 °C and 1 μl of a 1:50 dilution (approximately 0.6 units) of RNase-free DNase I (Roche) was added, and incubation was allowed to continue for 1 min. The reaction was terminated by addition of buffer T (250 μl of 10 mM Tris/pH 8.0, 10 mM EDTA, 750 mM NaCl, 1% SDS and 0.04 μg/μl glycogen) and the samples were purified using phenol/chloroform extraction and ethanol precipitation. After resuspension in formamide loading buffer, the samples were analyzed on a 6% denaturing polyacrylamide gel along with a sequence ladder that was generated using the same radiolabeled oligonucleotides.

ACKNOWLEDGEMENTS

We thank N.J. Robinson (Medical School, Newcastle University, Newcastle, UK) for helpful suggestions and critically reading the manuscript prior to submission and S.D. Bell (Medical Research Council Cancer Cell Unit, Hutchison MRC Research Centre, Cambridge, UK) for supplying *Sulfolobus* transcription factors and experimental advice.

REFERENCES

1 Aravind, L. and Koonin, E.V. (1999) DNA-binding proteins and evolution of transcription regulation in the archaea, *Nucleic Acids Res* 27, 4658-70

2 Bell, S.D. (2005) Archaeal transcriptional regulation - variation on a bacterial theme? *Trends Microbiol* 13, 262-5

3 Bell, S.D., Cairns, S.S., Robson, R.L. and Jackson, S.P. (1999) Transcriptional regulation of an archaeal operon *in vivo* and *in vitro*, *Mol Cell* 4, 971-82

4 Bell, S.D. and Jackson, S.P. (2000) Mechanism of autoregulation by an archaeal transcriptional repressor, *J Biol Chem* 275, 31624-9

5 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**, 38160-9

6 Schelert, J., Dixit, V., Hoang, V., Simbahan, J., Drozda, M. and Blum, P. (2004) Occurrence and characterization of mercury resistance in the hyperthermophilic archaeon *Sulfolobus solfataricus* by use of gene disruption, *J Bacteriol* **186**, 427-37

7 Vierke, G., Engelmann, A., Hebbeln, C. and Thomm, M. (2003) A novel archaeal transcriptional regulator of heat shock response, *J Biol Chem* **278**, 18-26

8 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**, 29537-49

9 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**, 5097-102

10 Gregor, D. and Pfeifer, F. (2001) Use of a halobacterial bgaH reporter gene to analyse the regulation of gene expression in halophilic archaea, *Microbiology* **147**, 1745-54

11 Hochheimer, A., Hedderich, R. and Thauer, R.K. (1999) The DNA binding protein Tfx from *Methanobacterium thermoautotrophicum*: structure, DNA binding properties and transcriptional regulation, *Mol Microbiol* **31**, 641-50

12 Ettema, T.J., Huynen, M.A., de Vos, W.M. and van der Oost, J. (2003) TRASH: a novel metal-binding domain predicted to be involved in heavy-metal sensing, trafficking and resistance, *Trends Biochem Sci* **28**, 170-3

13 Edwards, K.J., Bond, P.L., Gihring, T.M. and Banfield, J.F. (2000) An archaeal iron-oxidizing extreme acidophile important in acid mine drainage, *Science* **287**, 1796-9

14 Degtyarenko, K. (2000) Bioinorganic motifs: towards functional classification of metalloproteins, *Bioinformatics* **16**, 851-64

15 Nies, D.H. (1999) Microbial heavy-metal resistance, *Appl Microbiol Biotechnol* **51**, 730-50

16 Camakaris, J., Voskoboinik, I. and Mercer, J.F. (1999) Molecular mechanisms of copper homeostasis, *Biochem Biophys Res Commun* **261**, 225-32

17 Blindauer, C.A., Harrison, M.D., Robinson, A.K., Parkinson, J.A., Bowness, P.W., Sadler, P.J. and Robinson, N.J. (2002) Multiple bacteria encode metallothioneins and SmtA-like zinc fingers, *Mol Microbiol* **45**, 1421-32

18 Lutsenko, S. and Kaplan, J.H. (1996) P-type ATPases, *Trends Biochem Sci* **21**, 467

19 Rosenzweig, A.C. (2002) Metallochaperones: bind and deliver, *Chem Biol* **9**, 673-7

20 Harrison, M.D., Jones, C.E., Solioz, M. and Dameron, C.T. (2000) Intracellular copper routing: the role of copper chaperones, *Trends Biochem Sci* **25**, 29-32

21 Mercer, J.F. (2001) The molecular basis of copper-transport diseases, *Trends Mol Med* **7**, 64-9

22 Busenlehner, L.S., Pennella, M.A. and Giedroc, D.P. (2003) The SmtB/ArsR family of metalloregulatory transcriptional repressors: Structural insights into prokaryotic metal resistance, *FEMS Microbiol Rev* **27**, 131-43

23 Guedon, E. and Helmann, J.D. (2003) Origins of metal ion selectivity in the DtxR/MntR family of metalloregulators, *Mol Microbiol* **48**, 495-506

24 Brown, N.L., Stoyanov, J.V., Kidd, S.P. and Hobman, J.L. (2003) The MerR family of transcriptional regulators, *FEMS Microbiol Rev* **27**, 145-63

25 Cavet, J.S., Borrelly, G.P. and Robinson, N.J. (2003) Zn, Cu and Co in cyanobacteria: selective control of metal availability, *FEMS Microbiol Rev* 27, 165-81

26 Beck, C.F. and Warren, R.A. (1988) Divergent promoters, a common form of gene organization, *Microbiol Rev* 52, 318-26

27 Schultz, J., Milpetz, F., Bork, P. and Ponting, C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains, *Proc Natl Acad Sci U S A* 95, 5857-64

28 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, 287-94

29 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, 37464-8

30 Grishin, N.V. (2001) Treble clef finger--a functionally diverse zinc-binding structural motif, *Nucleic Acids Res* 29, 1703-14

31 Odermatt, A., Suter, H., Krapf, R. and Solioz, M. (1993) Primary structure of two P-type ATPases involved in copper homeostasis in *Enterococcus hirae*, *J Biol Chem* 268, 12775-9

32 Rensing, C., Fan, B., Sharma, R., Mitra, B. and Rosen, B.P. (2000) CopA: An *Escherichia coli* Cu(I)-translocating P-type ATPase, *Proc Natl Acad Sci U S A* 97, 652-6

33 Vats, N. and Lee, S.F. (2001) Characterization of a copper-transport operon, *cop*YAZ, from *Streptococcus mutans*, *Microbiology* 147, 653-62

34 Bayle, D., Wangler, S., Weitzenegger, T., Steinhilber, W., Volz, J., Przybylski, M., Schafer, K.P., Sachs, G. and Melchers, K. (1998) Properties of the P-type ATPases encoded by the *cop*AP operons of *Helicobacter pylori* and *Helicobacter felis*, *J Bacteriol* 180, 317-29

35 Bull, P.C., Thomas, G.R., Rommens, J.M., Forbes, J.R. and Cox, D.W. (1993) The Wilson disease gene is a putative copper transporting P-type ATPase similar to the Menkes gene, *Nat Genet* 5, 327-37

36 Hung, I.H., Casareno, R.L., Labesse, G., Mathews, F.S. and Gitlin, J.D. (1998) HAH1 is a copper-binding protein with distinct amino acid residues mediating copper homeostasis and antioxidant defense, *J Biol Chem* 273, 1749-54

37 Hamza, I., Schaefer, M., Klomp, L.W. and Gitlin, J.D. (1999) Interaction of the copper chaperone HAH1 with the Wilson disease protein is essential for copper homeostasis, *Proc Natl Acad Sci U S A* 96, 13363-8

38 Walker, J.M., Tsivkovskii, R. and Lutsenko, S. (2002) Metallochaperone Atox1 transfers copper to the NH2-terminal domain of the Wilson's disease protein and regulates its catalytic activity, *J Biol Chem* 277, 27953-9

39 Rost, B., Schneider, R. and Sander, C. (1997) Protein fold recognition by prediction-based threading, *J Mol Biol* 270, 471-80

40 Grogan, D.W. (1989) Phenotypic characterization of the archaebacterial genus *Sulfolobus*: comparison of five wild-type strains, *J Bacteriol* 171, 6710-9

41 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, 13662-7

42 Stoyanov, J.V., Hobman, J.L. and Brown, N.L. (2001) CueR (YbbI) of *Escherichia coli* is a MerR family regulator controlling expression of the copper exporter CopA, *Mol Microbiol* 39, 502-11

43 Outten, F.W., Outten, C.E., Hale, J. and O'Halloran, T.V. (2000) Transcriptional activation of an *Escherichia coli* copper efflux regulon by the chromosomal MerR homologue, *cue*R, *J Biol Chem* 275, 31024-9

44 Strausak, D. and Solioz, M. (1997) CopY is a copper-inducible repressor of the Enterococcus hirae copper ATPases, *J Biol Chem* **272**, 8932-6

45 Carey, J. (1991) Gel retardation, *Methods Enzymol* **208**, 103-17

46 Dixit, V., Bini, E., Drozda, M. and Blum, P. (2004) Mercury inactivates transcription and the generalized transcription factor TFB in the archaeon *Sulfolobus solfataricus*, *Antimicrob Agents Chemother* **48**, 1993-9

47 Deigweiher, K., Drell, T.L.t., Prutsch, A., Scheidig, A.J. and Lubben, M. (2004) Expression, isolation, and crystallization of the catalytic domain of CopB, a putative copper transporting ATPase from the thermoacidophilic archaeon *Sulfolobus solfataricus*, *J Bioenerg Biomembr* **36**, 151-9

48 Multhaup, G., Strausak, D., Bissig, K.D. and Solioz, M. (2001) Interaction of the CopZ copper chaperone with the CopA copper ATPase of *Enterococcus hirae* assessed by surface plasmon resonance, *Biochem Biophys Res Commun* **288**, 172-7

49 Tottey, S., Rondet, S.A., Borrelly, G.P., Robinson, P.J., Rich, P.R. and Robinson, N.J. (2002) A copper metallochaperone for photosynthesis and respiration reveals metal-specific targets, interaction with an importer, and alternative sites for copper acquisition, *J Biol Chem* **277**, 5490-7

50 Cobine, P., Wickramasinghe, W.A., Harrison, M.D., Weber, T., Solioz, M. and Dameron, C.T. (1999) The Enterococcus hirae copper chaperone CopZ delivers copper(I) to the CopY repressor, *FEBS Lett* **445**, 27-30

51 Cobine, P.A., George, G.N., Jones, C.E., Wickramasinghe, W.A., Solioz, M. and Dameron, C.T. (2002) Copper transfer from the Cu(I) chaperone, CopZ, to the repressor, Zn(II)CopY: metal coordination environments and protein interactions, *Biochemistry* **41**, 5822-9

52 Cobine, P.A., Jones, C.E. and Dameron, C.T. (2002) Role for zinc(II) in the copper(I) regulated protein CopY, *J Inorg Biochem* **88**, 192-6

53 Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* **22**, 4673-80

54 Letunic, I., Copley, R.R., Schmidt, S., Ciccarelli, F.D., Doerks, T., Schultz, J., Ponting, C.P. and Bork, P. (2004) SMART 4.0: towards genomic data integration, *Nucleic Acids Res* **32 Database issue**, D142-4

55 Schwede, T., Kopp, J., Guex, N. and Peitsch, M.C. (2003) SWISS-MODEL: An automated protein homology-modeling server, *Nucleic Acids Res* **31**, 3381-5

56 Ban, N., Nissen, P., Hansen, J., Moore, P.B. and Steitz, T.A. (2000) The complete atomic structure of the large ribosomal subunit at 2.4 A resolution, *Science* **289**, 905-20

57 Guex, N. and Peitsch, M.C. (1997) SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling, *Electrophoresis* **18**, 2714-23

58 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**, 529-40

Chapter 6

A novel ligand-binding domain involved in regulation of amino acid metabolism in prokaryotes

Thijs J. G. Ettema
Arie B. Brinkman
Travis H. Tani
John B. Rafferty
John van der Oost

A combination of sequence profile searching and structural protein analysis has revealed a novel type of small molecule binding domain that is involved in the allosteric regulation of prokaryotic amino acid metabolism. This domain, designated RAM, has been found to be fused to the DNA-binding domain of Lrp-like transcription regulators and to the catalytic domain of some metabolic enzymes, and has been found as a stand-alone module. Structural analysis of the RAM domain of Lrp reveals a βαββαβ-fold that is strikingly similar to that of the recently described ACT domain, a ubiquitous allosteric regulatory domain of many metabolic enzymes. However, structural alignment and re-evaluation of previous mutagenesis data suggest that the effector-binding sites of both modules are significantly different. By assuming that the RAM and ACT domains originated from a common ancestor, these observations suggest that their ligand-binding sites have evolved independently. Both domains appear to play analogous roles in controlling key steps in amino acid metabolism at the level of gene expression as well as enzyme activity.

Journal of Biological Chemistry 277(40), 37464-37468 (2002)

INTRODUCTION

Allosteric regulation is a general mechanism that enables a tight control of both enzyme activity and gene expression. A textbook example of this mechanism is the feedback inhibition of enzymes that catalyze key steps in amino acid biosynthesis. Important insight in the molecular basis of modulated enzyme activity has been provided by the crystal structure of 3-phosphoglycerate dehydrogenase (SerA)[1,2], an enzyme that catalyzes the rate-limiting first step in serine biosynthesis. The SerA structure has revealed separated catalytic and regulatory domains that are connected by a flexible hinge. The structural motif of the regulatory domain consists of a four-stranded anti-parallel -sheet with two -helices packed on one side. When the serine effector molecule binds to this αβ-sandwich, the hinge region allows allosteric regulation: a slight inter-domain rearrangement that down-regulates the catalytic activity of the enzyme[2,3].

A thorough sequence profile analysis has shown that the SerA regulatory domain is an ancient small molecule binding domain (SMBD)[4] that is conserved in a wide variety of enzymes as well as in some transcriptional regulators that are involved in the control of amino acid and purine metabolism[5]. As predicted in the latter study, the presence of this ancient "ACT domain" (for review see reference 6) was indeed demonstrated in the recent crystal structures of rat phenylalanine hydroxylase[7], an enzyme that catalyzes the conversion of phenylalanine to tyrosine. Interestingly, the structure of the *Escherichia coli* threonine deaminase[8], the enzyme catalyzing the first step of the isoleucine biosynthesis pathway, revealed two domains that resembled the ACT domain of SerA at the structural level rather than at the sequence level[5]. This observation demonstrates that the sequence divergence of SMBDs like ACT can expand beyond the detection limits of the sequence-based algorithm of PSI-BLAST[9]. For this reason, the threonine deaminase-ACT domains have been referred to as "ACT-like" domains[6].

In the present study we describe a novel ligand-binding module that we named the RAM domain because of its general involvement in the allosteric Regulation of Amino acid Metabolism. This domain is mainly found in association with a class of prokaryotic transcriptional regulators but also as a module in enzymes and in some instances as stand-alone SMBD.

RESULTS

The structure of the C-terminal domain of LrpA resembles the structure of the ACT domain

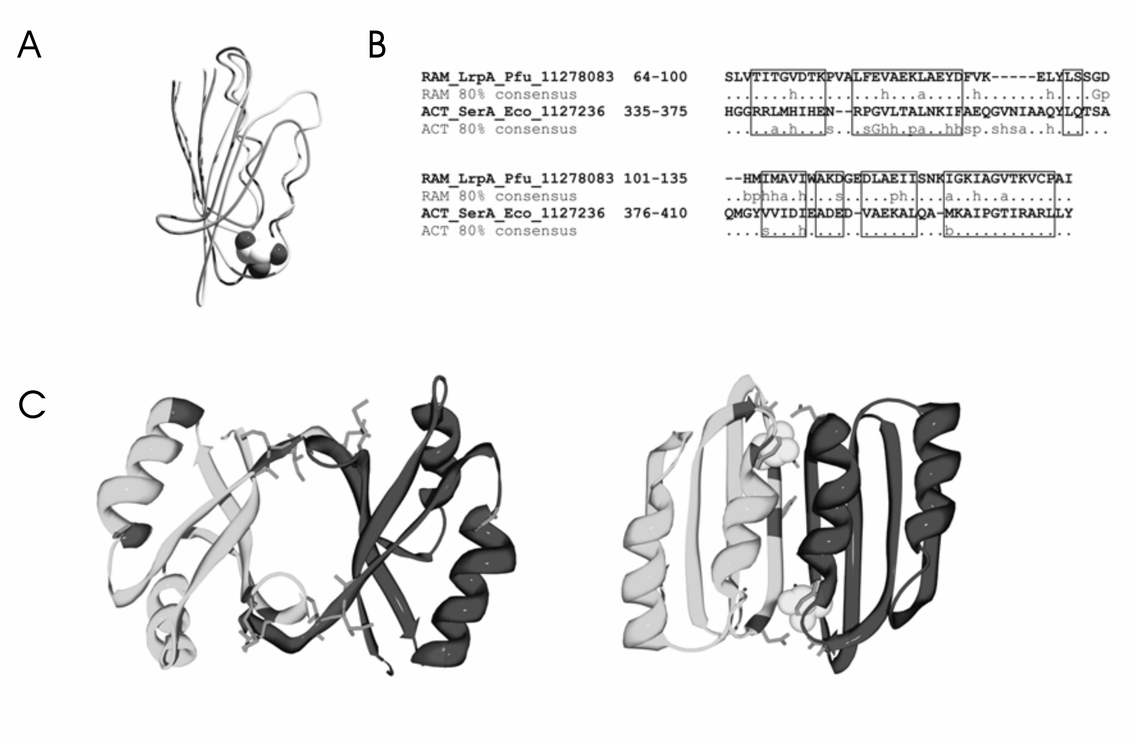

A
B
RAM_LrpA_Pfu_11278083 64-100 SLVTITGVDTKPVALFEVAEKLAEYDFVK-----ELYLSSGD
RAM 80% consensus ....h.........h...a.....h........h...Gp
ACT_SerA_Eco_1127236 335-375 HGGRRLMHIHEN--RPGVLTALNKIFAEQGVNIAAQYLQTSA
ACT 80% consensus .....a.h...s...sGhh.pa..hhsp.shsa..h.......
RAM_LrpA_Pfu_11278083 101-135 --HMIMAVIWAKDGEDLAEIISNKIGKIAGVTKVCPAI
RAM 80% consensus ..bphha.h...s.....ph....a..h..a.......
ACT_SerA_Eco_1127236 376-410 QMGYVVIDIEADED-VAEKALQA-MKAIPGTIRARLLY
ACT 80% consensus ....s...h...............b.............
C

Figure 6.1 (NB: For COLOR version of this figure, see APPEDIX I) (a) Superimposition of the RAM domain of the *P. furiosus* LrpA (*purple*) with the ACT domain of the *E. coli* SerA (*yellow*) showing the strong similarity between the two domains at the structural level (root mean square deviation value 1.8 Å). The superimposition was constructed with 49 -carbon atoms, which composes about 65% of the domains. In addition, the position of the negative effector of SerA ACT domain, serine, was indicated within the superimposed domains. Superimposition and figure were created using the Swiss PDB viewer[10]. **(b)** Structural alignment of the RAM domain of the *P. furiosus* LrpA (residues 64-135) and the ACT domain of the *E. coli* SerA (residues 335-410). In addition, the 80% consensus sequences for RAM and ACT domains are included in the alignment indicating the sequence divergence between the two SMBDs. For abbreviations of the different amino acid classes see figure 2. Matched residues that display a root mean square value of less than 2.5 Å are boxed. The structural alignment was constructed using the Swiss PDB viewer[10] using the "Structural alignment" option. **(c)** Structural comparison of the RAM dimer of *P. furiosus* LrpA (*left*) and ACT dimer of *E. coli* SerA (*right*). The monomers are shown in *cyan* and *blue*, and the ligand response mutations of the RAM domain and the ACT domain corresponding to those that are depicted in figure 2, (a) and (b), respectively, were mapped into the backbones of the respective structures in *red* with *magenta* side chains.

The Lrp family of transcriptional regulators plays a crucial role in the control of amino acid metabolism in prokaryotes. Although the leucine-responsive regulatory protein (Lrp) from *E. coli* is a global transcriptional regulator[11], most Lrp homologs act as specific regulators (e.g. refs. 12-14). The interaction of specific amino acid effectors with Lrp-like regulators may lead to modulation of (i) DNA affinity, (ii) DNA bending, (iii) Lrp oligomeric state (dimer/tetramer/octamer/hexadecamer), and (iv) Lrp tertiary structure[6,12,13,15]. All these changes most likely reflect an allosteric regulation of Lrp activity by a (minor) structural rearrangement.

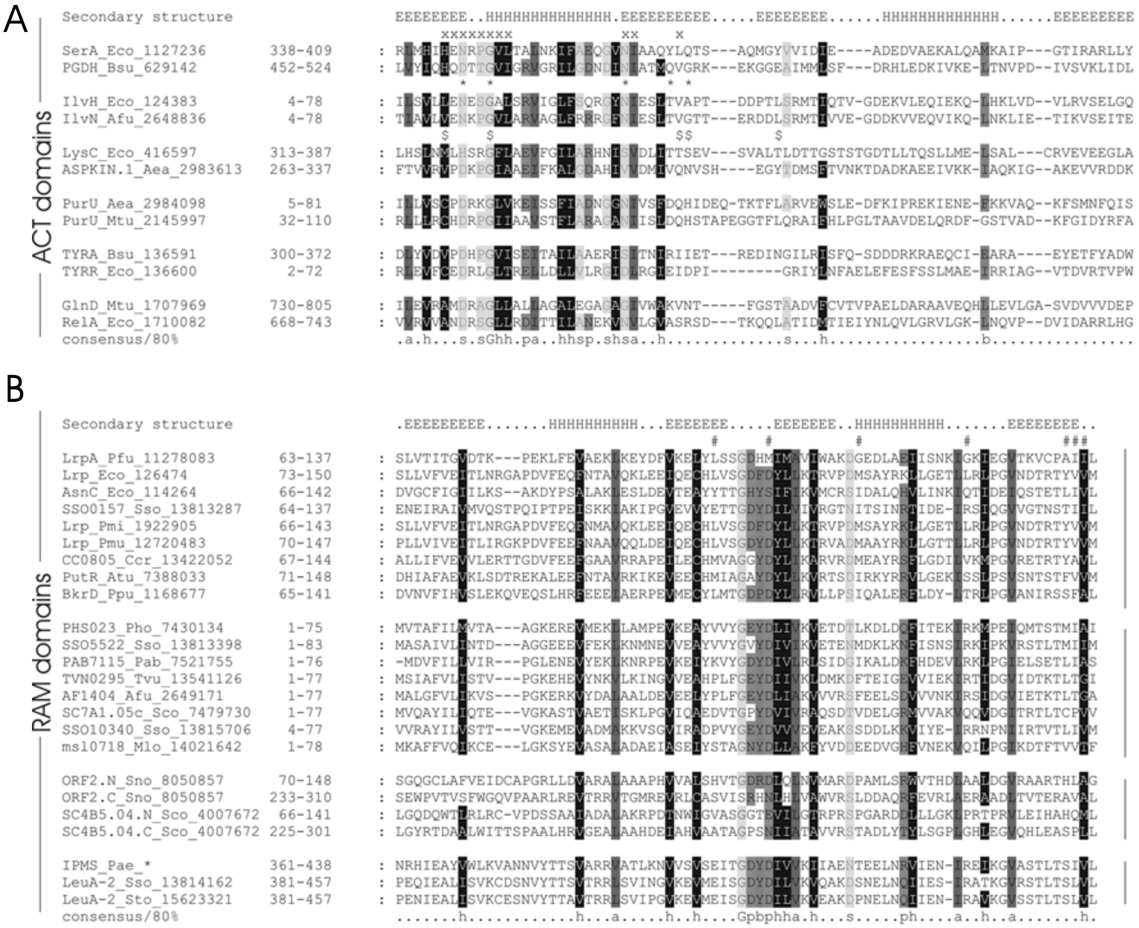


Figure 6.2 (NB: For COLOR version of this figure, see APPEDIX I) (a) Alignment of the most diverse members of the ACT domain. The secondary structure assignment that is indicated above the alignment was derived from the crystal structure of the *E. coli* 3-phosphoglycerate dehydrogenase (SerA; PDB code 1PSD). The amino acid residues that are involved in the binding of the effector serine in SerA, are indicated above the alignment (x). In addition, the ligand response mutations that were determined for the *E. coli* small subunit of the acetolactate synthase (IlvH) and aspartokinase (LysC) are indicated with * and $, respectively. The 80% consensus shown below the alignments was obtained as described above, and the position numbers on the left side indicate the limits of the domains. Also a structural alignment of the two regulatory ACT-like domains of the *E. coli* threonine deaminase (THD1) is included, with a mapped ligand response mutation (+), indicating the divergence on the sequence level between these domains and the genuine ACT domains. Secondary structure from the first repeat of THD1 is indicated above the alignment. The 80% consensus shown below the alignments was obtained using the following amino acid classes (4): small (*s*, ACGSTDNVP; *shaded yellow*), polar (*p*, YWHKREQDNST; *shaded green*), big (*b*, FILMVWYKREQ; *gray shading*), and hydrophobic (*h*, ILVCAGMFYWHTP; *shaded black*). Of the hydrophobic residues, the aliphatic subset (h, ILVA) is in black with red shading. The position numbers

Figure 6.2 (continued…)
on the *left side* indicate the limits of the domains. **(b)** Multiple alignment of RAM domains. By using the structure of *P. furiosus* LrpA[16], the secondary structure of the RAM domain has been deduced, as is indicated *above* the alignment (*E*, β-strand; *H*, α-helix). The sequences are grouped by domain architecture as follows: 1, Lrp-like transcriptional regulators and homologs (HTH-RAM); 2, stand-alone versions of the RAM domain; 3, duplication of Lrp-like transcriptional regulator; and 4, RAM domain present in crenarchaeal 2-isopropylmalate synthase. The ligand response mutations as determined for the *E. coli* Lrp by Platko and Calvo[17] are indicated abovethe alignment (#). All protein sequences were obtained from ENTREZ, and GenBank™ identifiers are indicated for each protein. Species abbreviations are as follows: Aea: *Aquifex aeoliticus*; Afu: *Archaeoglobus fulgidus*; Atu: *Agrobacterium tumefaciens*; Bsu: *Bacillus subtilis*; Ccr: *Caulobacter crescentus*; Eco: *E. coli*; Mlo: *Mesorhizobium loti*; Mtu: *Mycobacterium tuberculosis*; Pab: *Pyrococcus abyssi*; Pfu: *P. furiosus*; Pho: *Pyrococcus horikoshii*; Pmi: *Proteus mirabilis*; Pmu: *Pasteurella multocida*; Ppu: *Pseudomonas putida*; Sco: *Streptomyces coelicolor*; Sno: *Streptomyces noursei*; Sso: *S. solfataricus*; Sto: *S. tokodaii* and Tvu: *Thermoplasma vulcanium*.

Recently, the structure of an archaeal Lrp homolog, the *P. furiosus* LrpA octamer (or tetramer of dimers), has been resolved[16]. The structure of the LrpA monomer revealed an N-terminal DNA-binding helix-turn-helix (HTH) domain, an extended hinge, and a C-terminal globular domain with a βαββαβ-fold. Previous mutagenesis analyses were in perfect agreement with the N-terminal domain being involved in the interaction with DNA and predicted the C-terminal αβ-sandwich to have a regulatory function[16,17]. What was not noted during the initial analysis of the LrpA structure is the interesting fact that the C-terminal regulatory domain of LrpA appears to resemble the ACT domains[5,6] with respect to structure and function; both consist of a typical αβ-sandwich and are anticipated to be regulatory domains involved in allosteric modulation of the activity of enzymes and DNA-binding proteins that are involved in amino acid metabolism. Despite the similarity in structure with the ACT domain, we propose that, for reasons discussed below, the C-terminal regulatory domain of LrpA is part of a novel, distinct class of regulatory domains.

The overall structural resemblance between the RAM domain of the *P. furiosus* LrpA and the ACT domain of the *E. coli* SerA is confirmed by superimposition of both structures (Figure 6.1a). The obtained root mean square deviation value between the LrpA-RAM and the SerA-ACT (1.8 Å) compares with a value obtained with a superimposition between SerA-ACT and the ACT domain of the rat phenylalanine hydroxylase (1.7 Å).

Ligand response mutations suggest a different location of the ligand-binding sites in RAM and ACT

Despite the structural similarities between the RAM domain and the ACT domain, however, the effector-binding sites in these domains seem to be different. In the ACT domain of SerA, the loop that links the first β-strand (β1) and the first α-helix (α1) of the βαββαβ-fold, makes up the binding pocket of the serine effector. The important role of this loop is in agreement with the fact that it is very well conserved in ACT-containing enzymes and regulators (Figure 6.2a). An invariant glycine residue, which is also conserved in the ACT-like domains of the *E. coli* threonine deaminase, and an adjacent hydrophobic residue have been proposed to be involved in maintenance of the strand-helix interface; two additional conserved polar residues are involved in binding the ligand with hydrogen bonds[5]. The importance of the conserved region was also confirmed by mutation analysis of other ACT-containing enzymes from *E. coli*, i.e. the valine-binding regulatory subunit of the acetolactate synthase (IlvH)[18] and the lysine-sensitive aspartokinase (LysC)[19] (Figure 6.2a). Mutations that resulted in a strongly reduced or abolished response after being exposed to their respective effector (ligand response mutations) all cluster within the conserved loop region, suggesting that binding of the effector resembles the interaction of SerA with serine. Mapping the ACT ligand response mutations into the structure of the ACT dimer of SerA (Figure 6.1c) confirms this idea. The ligand response mutations cluster at the dimer interface in general and at the ligand (serine)-binding site in particular (Figure 6.1c).

The dimer structure of the SerA-ACT differs significantly from the LrpA-RAM dimer. Whereas the contact between the two ACT domains seems to be mediated via the α2 and β3 interface resulting in an eight-stranded anti-parallel β-sheet[1], in LrpA the RAM dimer is mainly formed by interactions between the antiparallel β-sheets that are facing each other, forming an antiparallel β-barrel-like structure (Figure 6.1c)[16]. Because of this structural difference, and given the fact that the ligand-binding site for ACT domains is located at the dimer interface, this might imply that the ligand-binding site is different in RAM domains. Indeed, the region that is involved in interaction with the ligand (Figure 6.2a) is well conserved in ACT domains, whereas the corresponding loop that links strand β1 and helix α1 in RAM domains displays only poor sequence conservation (Figure 6.2b). In RAM, the best conserved region, again including an invariant glycine residue, appears to be the region surrounding the loop connecting strands β2 and β3. Moreover, extensive mutagenesis studies that have been performed with the *E. coli* Lrp[17] confirm the importance of this region with respect to ligand response. The Lrp leucine response mutations that were obtained in this study apparently lost the capacity to bind their ligand[17]. When the equivalents of these *E. coli* Lrp ligand response mutations are mapped into the structure of the LrpA-RAM dimer of *P. furiosus*, it becomes clear that five of seven mapped mutations belonging to this class (Leu-95, Met-101, Ala-134, Ile-135, and Ile-136) are clustered in a region across the dimer interface (Figure 6.1c). The remaining two ligand response mutations (Gly-111 and Gly-123) are located in close proximity to the other mutations, albeit in adjacent dimers of the octamer rather than within the same dimer[16] (not shown in Figure 6.1c). These observations suggest that the ligand-binding site of RAM is located at a different position than that of ACT. Based on the relatively high sequence conservation and to some extent on the mutation data, it is tempting to speculate that the ligand-binding site of RAM is located between the β2 and β3. Lrp ligand co-crystallization experiments are required to confirm this hypothesis.

DISCUSSION

The βαββαβ-motif appears to be a common regulatory structure in amino acid metabolic enzymes and transcriptional regulators; both the RAM and the ACT domains share this fold and are associated with proteins that are involved with amino acid metabolism either as part of enzymes, as part of transcriptional regulators, or as stand-alone SMBD. Apart from the structural and functional similarity between the two domains, another connection is the fact that the expression of some bacterial ACT-containing enzymes (e.g. SerA and IlvHI) is under control of RAM-containing transcriptional regulators of the Lrp family[11]. These observed analogies between RAM and ACT may be useful for speculating about the function of uncharacterized RAM and ACT domains. For example, the function of the stand-alone versions of the RAM domains that are present in several bacterial and archaeal genomes is yet unclear. However, the function of stand-alone ACT domains might suggest the possible function of their stand-alone RAM counterparts. For example, the acetolactate synthase in *E. coli* is a key enzyme in branched chain amino acid biosynthesis that is subjected to valine feedback inhibition[17]. The heterotetrameric holoenzyme is made up of the large catalytic subunit (IlvI) and a small regulatory subunit that consists of a single ACT domain (IlvH) and accounts for the valine-mediated feedback repression. It is possible that a stand-alone RAM domain performs a function that is analogous to that of the IlvH subunit, allosteric regulation of enzymes (or possibly transcriptional regulators) involved in amino acid metabolism via protein-protein interactions. Interestingly, profile-based analysis of prokaryal genomes (e.g. *S. solfataricus*) failed to identify a gene encoding the small regulatory subunit of the acetolactate synthase (IlvH), whereas homologs of the gene encoding the large catalytic subunit of the acetolactate synthase could be identified on the genome (e.g. ilvB-1). Possibly, the large subunit of the acetolactate synthase present in these organisms is subjected to allosteric regulation by the stand-alone RAM domains that are encoded on the genome.

A more clear-cut example is the anticipated role for the RAM domain that is present at the C terminus of the crenarchaeal IPMS. Generally, IPMS catalyzes the first step in leucine biosynthesis that is subjected to leucine-mediated feedback inhibition. Leucine has the following two known effects on this enzyme in *Salmonella typhimurium*: (i) it controls the catalytic activity of the enzyme by feedback inhibition[20], and (ii) it causes a dissociation of the tetrameric enzyme into its monomeric subunits[21]. Genetic mapping of leucine-insensitive mutants[22,23] revealed that the mutation resulting in the affected end product inhibition is located at the C terminus of the *S. typhimurium* IPMS. So the domain that is probably responsible for the allosteric control of IPMS is located in the C-terminal part of the enzyme. Sequence analysis of the C-terminal part of the *S. typhimurium* IPMS did not reveal the presence of a RAM or ACT domain, possibly suggesting the presence of yet another alternative regulatory domain. Although the exact mechanism of inhibition remains to be elucidated, the presence of a C-terminal RAM domain in the crenarchaeal RAM-IPMS strongly suggests that these enzymes are subjected to RAM-mediated feedback regulation.

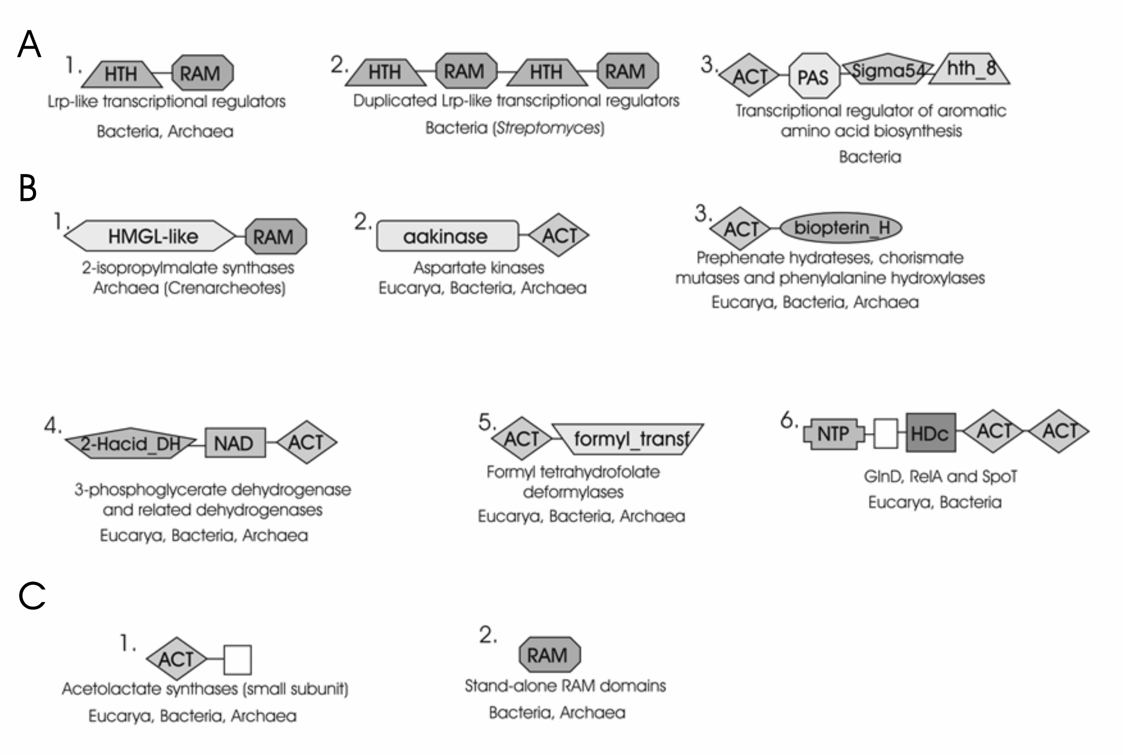

A
1. HTH RAM
Lrp-like transcriptional regulators
Bacteria, Archaea
2. HTH RAM HTH RAM
Duplicated Lrp-like transcriptional regulators
Bacteria (Streptomyces)
3. ACT PAS Sigma54 hth_8
Transcriptional regulator of aromatic amino acid biosynthesis
Bacteria
B
1. HMGL-like RAM
2-isopropylmalate synthases
Archaea (Crenarcheotes)
2. aakinase ACT
Aspartate kinases
Eucarya, Bacteria, Archaea
3. ACT biopterin_H
Prephenate hydrateses, chorismate mutases and phenylalanine hydroxylases
Eucarya, Bacteria, Archaea
4. 2-Hacid_DH NAD ACT
3-phosphoglycerate dehydrogenase and related dehydrogenases
Eucarya, Bacteria, Archaea
5. ACT formyl_transf
Formyl tetrahydrofolate deformylases
Eucarya, Bacteria, Archaea
6. NTP HDc ACT ACT
GlnD, RelA and SpoT
Eucarya, Bacteria
C
1. ACT
Acetolactate synthases (small subunit)
Eucarya, Bacteria, Archaea
2. RAM
Stand-alone RAM domains
Bacteria, Archaea

Figure 6.3 (previous page...) Domain architectures of ACT and RAM containing proteins, subdivided into three classes according to their function. For each class of proteins, the phyletic distribution is depicted below the domain structure. Domain analysis was performed using SMART[24]. Apart from the regulatory domains (ACT or RAM), the other domains that are part of the proteins are described as follows. **(a)** Transcription regulation-associated regulatory domains. *1*, Lrp-like transcriptional regulators, generally consisting of an N-terminal DNA-binding helix-turn-helix domain fused to a C-terminal RAM domain; *2*, duplicated form of Lrp-like transcriptional regulator, consisting of a tandem repeat of the transcriptional regulator; *3*, transcriptional regulator of aromatic amino acid biosynthesis, containing an N-terminal ACT domain, followed by a PAS domain which is possibly involved in signal sensing. The C-terminal part of these regulators contain a sigma54 interacting domain (PF00989) and a DNA-binding helix-turn-helix (hth_8; PF02954). **(b)** Enzyme-associated regulatory domains. *1*, Crenarchaeal 2-isopropylmalate synthases, containing an HMGL-like domain, which is found in a diverse set of enzymes including several aldolases and a pyruvate carboxylase; *2*, aspartokinases, containing an N-terminal kinase domain (PF00696) that is involved in phosphorylation of a variety of amino acid substrates; *3*, diverse group of proteins involved in biosynthesis of aromatic amino acids consisting of prephenate hydratases, chorismate mutases, and phenylalanine hydroxylases. From the latter enzyme the domain architecture is displayed, containing a biopterin_H domain (PF00351), which is present in biopterin-dependent aromatic amino acid hydroxylases; *4*, group of proteins consisting of 3-phosphoglycerate, homoserine, and malate dehydrogenases, containing an N-terminal 2-Hacid_DH catalytic domain (PF00389) and a NAD-binding domain (2-Hacid_DH_C, PF02826); *5*, formyltetrahydrofolate deformylases, typically containing a C-terminal formyl_transf. domain (PF00551), a domain that is present in multiple enzymes that are involved in de novo purine biosynthesis; *6*, diverse set of proteins containing uridylyltransferases that are involved in glutamine synthase regulation (GlnD), guanosine polyphosphate 3'-pyrophosphorylases (SpoT), and GTP pyrophosphokinases (RelA). The latter two enzymes are involved in stringent response. The domain architecture that is depicted here represents GlnD, containing a nucleotidyltransferase domain (NTP_transf, PF01909) and an HDc domain that is involved in metal-dependent phosphohydrolase activity. **(c)** Stand-alone regulatory domains. *1*, small regulatory domain of the acetolactate synthase, generally consisting of an N-terminal ACT domain fused to a small domain that is probably involved in the interaction with the large subunit (IlvI); *2*, isolated RAM domains. The function of these proteins is still to be elucidated; however, it is possible that they play a role analogous to the isolated ACT domains, i.e. regulatory subunit of enzymes or transcriptional regulators.

The date presented here indicate that structure-based alignments of the regulatory domains as well as extensive (reverse) PSI-BLAST analyses fail to close the apparent gap in sequence divergence between RAM and ACT. This might suggest that ACT and RAM both have independently evolved into allosteric regulatory domains. However, the most parsimonious scenario is that the two domains have emerged from a common ancestor and only evolved different interaction specificity with respect to their ligands. Possibly, the diversification in binding sites is the underlying reason of the dramatic sequence divergence between RAM and ACT, with the overall structure being well conserved. This appears to be yet another example of domain evolution in which sequence similarity has been lost while retaining their structural similarity[4].

EXPERIMENTAL PROCEDURES

PSI-BLAST analysis and multiple sequence alignments, domain analysis of proteins

In order to verify and characterize the relationship between distant RAM domains at the sequence level, we performed several PSI-BLAST searches[9] at the National Center of Biotechnology Information. When a PSI-BLAST search was seeded with the C-terminal domain of the *Pyrococcus furiosus* LrpA (residues 62-141), using a BLOSUM80 matrix and an expect value threshold of 0.001, the first stand-alone versions of RAM (lacking the HTH domain) were retrieved within the first iteration (10580664; $E = 6 \times 10^{12}$); the RAM domains within the *Sulfolobus solfataricus* and *Sulfolobus tokodaii* 2-isopropylmalate synthase (13814162, respectively, 15623321) were recovered in iteration 3 ($E = 7 \times 10^{6}$). A reverse PSI-BLAST using the *S. solfataricus* 2-isopropylmalate synthase sequence (residues 342-461) recovered HTH-RAM proteins (e.g. 13813287 at iteration 1, $E = 3 \times 10^{5}$) and stand-alone versions of the RAM domain (e.g. 13813398 at iteration 2, $E = 5 \times 10^{5}$), thereby connecting the most distant RAM domain containing proteins with each other on a statistical basis. A multiple alignment of RAM domains was constructed by collecting the "highest scoring pairs of sequence segments" using PSI-BLAST, which were re-aligned using ClustalW[25], followed by minor manual adjustment based on the secondary structure. Domain analysis of RAM- and ACT-containing proteins was performed using SMART[24] and PFAM[26]. All protein sequences were extracted from ENTREZ.

Structural alignment and superimposition three-dimensional structures of SerA ACT domain and LrpA RAM domain

In order to detect the structural similarity of ACT and RAM, a superimposition of the C-terminal domain of the *P. furiosus* LrpA (residues 64-135) and the ACT domain present in the SerA of *E. coli* (residues 335-410, PDB entry 1PSD) was constructed using the Swiss PDB viewer[10] using the "Iterative fit" option. The superimposition was constructed with 49 -carbon atoms that displayed a root mean square value of less than 2.5 Å, comprising about 65% of the domains. From the superimposed structures, a structural alignment was deduced using the Structural alignment tool of the Swiss PDB viewer[10].

ACKNOWLEDGEMENTS

We thank Dr. Eugene V. Koonin (National Center of Biotechnology Information, Bethesda) for stimulating discussions.

REFERENCES

1 Schuller, D.J., Grant, G.A. and Banaszak, L.J. (1995) The allosteric ligand site in the V^{max}-type cooperative enzyme phosphoglycerate dehydrogenase, *Nat Struct Biol* **2**, 69-76

2 Grant, G.A., Schuller, D.J. and Banaszak, L.J. (1996) A model for the regulation of D-3-phosphoglycerate dehydrogenase, a V^{max}-type allosteric enzyme, *Protein Sci* **5**, 34-41

3 Al-Rabiee, R., Zhang, Y. and Grant, G.A. (1996) The mechanism of velocity modulated allosteric regulation in D-3-phosphoglycerate dehydrogenase. Site-directed mutagenesis of effector binding site residues, *J Biol Chem* **271**, 23235-8

4 Anantharaman, V., Koonin, E.V. and Aravind, L. (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains, *J Mol Biol* **307**, 1271-92

5 Aravind, L. and Koonin, E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches, *J Mol Biol* **287**, 1023-40

6 Chipman, D.M. and Shaanan, B. (2001) The ACT domain family, *Curr Opin Struct Biol* **11**, 694-700

7 Kobe, B., Jennings, I.G., House, C.M., Michell, B.J., Goodwill, K.E., Santarsiero, B.D., Stevens, R.C., Cotton, R.G. and Kemp, B.E. (1999) Structural basis of autoregulation of phenylalanine hydroxylase, *Nat Struct Biol* **6**, 442-8

8 Gallagher, D.T., Gilliland, G.L., Xiao, G., Zondlo, J., Fisher, K.E., Chinchilla, D. and Eisenstein, E. (1998) Structure and control of pyridoxal phosphate dependent allosteric threonine deaminase, *Structure* **6**, 465-75

9 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**, 3389-402

10 Schwede, T., Diemand, A., Guex, N. and Peitsch, M.C. (2000) Protein structure computing in the genomic era, *Res Microbiol* **151**, 107-12

11 Calvo, J.M. and Matthews, R.G. (1994) The leucine-responsive regulatory protein, a global regulator of metabolism in *Escherichia coli*, *Microbiol Rev* **58**, 466-90

12 Madhusudhan, K.T., Huang, N., Braswell, E.H. and Sokatch, J.R. (1997) Binding of L-branched-chain amino acids causes a conformational change in BkdR, *J Bacteriol* **179**, 276-9

13 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**, 29537-49

14 Kolling, R., Gielow, A., Seufert, W., Kucherer, C. and Messer, W. (1988) AsnC, a multifunctional regulator of genes located around the replication origin of E*scherichia coli*, *ori*C, *Mol Gen Genet* **212**, 99-104

15 Chen, S., Rosner, M.H. and Calvo, J.M. (2001) Leucine-regulated self-association of leucine-responsive regulatory protein (Lrp) from *Escherichia coli*, *J Mol Biol* **312**, 625-35

16 Leonard, P.M., Smits, S.H., Sedelnikova, S.E., Brinkman, A.B., de Vos, W.M., van der Oost, J., Rice, D.W. and Rafferty, J.B. (2001) Crystal structure of the Lrp-like transcriptional regulator from the archaeon *Pyrococcus furiosus*, *Embo J* **20**, 990-7

17 Platko, J.V. and Calvo, J.M. (1993) Mutations affecting the ability of *Escherichia coli* Lrp to bind DNA, activate transcription, or respond to leucine, *J Bacteriol* **175**, 1110-7

18 Mendel, S., Elkayam, T., Sella, C., Vinogradov, V., Vyazmensky, M., Chipman, D.M. and Barak, Z. (2001) Acetohydroxyacid synthase: a proposed structure for regulatory subunits supported by evidence from mutagenesis, *J Mol Biol* **307**, 465-77

19 Kikuchi, Y., Kojima, H. and Tanaka, T. (1999) Mutational analysis of the feedback sites of lysine-sensitive aspartokinase of *Escherichia coli*, *FEMS Microbiol Lett* **173**, 211-5

20 Kohlhaw, G., Leary, T.R. and Umbarger, H.E. (1969) Alpha-isopropylmalate synthase from *Salmonella typhimurium*. Purification and properties, *J Biol Chem* **244**, 2218-25

21 Leary, T.R. and Kohlhaw, G. (1970) Dissociation of alpha-isopropylmalate synthase from *Salmonella typhimurium* by its feedback inhibitor leucine, *Biochem Biophys Res Commun* **39**, 494-501

22 Calvo, R.A. and Calvo, J.M. (1967) Lack of end-product inhibition and repression of leucine synthesis in a strain of *Salmonella typhimurium*, *Science* **156**, 1107-9

23 Bartholomew, J.C. and Calvo, J.M. (1971) Alpha-isopropylmalate synthase from *Salmonella typhimurium*. Carboxypeptidase digestion studies of parent and feedback-insensitive enzymes, *Biochim Biophys Acta* **250**, 568-76

24 Schultz, J., Milpetz, F., Bork, P. and Ponting, C.P. (1998) SMART, a simple modular architecture research tool: identification of signaling domains, *Proc Natl Acad Sci U S A* **95**, 5857-64

25 Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* **22**, 4673-80

26 Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. and Sonnhammer, E.L. (2002) The Pfam protein families database, *Nucleic Acids Res* **30**, 276-80

Chapter 7

Identification and functional verification of archaeal-type phospho*enol*pyruvate carboxylase, a missing link in archaeal central carbohydrate metabolism

Thijs J. G. Ettema
Kira S. Makarova
Gera L. Jellema
Hinco J. Gierman
Eugene V. Koonin
Martijn A. Huynen
Willem M. de Vos
John van der Oost

Despite the fact that phosphoenolpyruvate carboxylase (PEPC) activity has been measured and in some cases even purified from some Archaea, the gene responsible for this activity has not been elucidated. Using sensitive sequence comparison methods, we detected a highly conserved, uncharacterized archaeal gene family that is distantly related to the catalytic core of the canonical PEPC. To verify the predicted function of this archaeal gene family, we cloned a representative from the hyperthermophilic acidophile *Sulfolobus solfataricus* and functionally produced the corresponding enzyme as a fusion with the *Escherichia coli* maltose-binding protein. The purified fusion protein indeed displayed highly thermostable PEPC activity. The structural and biochemical properties of the characterized archaeal-type PEPC (atPEPC) from *S. solfataricus* are in good agreement with previously reported biochemical analyses of other archaeal PEPC enzymes. The newly identified atPEPC, with its distinct properties, constitutes yet another example of the versatility of the enzymes of the central carbon metabolic pathways in the archaeal domain.

Journal of Bacteriology **186**(22), 7754-7762 (2004)

INTRODUCTION

Phosphoenolpyruvate carboxylase (PEPC, EC4.1.1.31) catalyzes the irreversible β-carboxylation of phosphoenolpyruvate (PEP) to form oxaloacetate (OAA) and inorganic phosphate using HCO_3^- as a co-substrate and divalent metal ions as cofactors. PEPC is a cytosolic enzyme that is widely distributed among a great variety of Bacteria, as well as higher and lower plants and performs diverse biological functions[1]. In particular, PEPC plays an important role in heterotrophic bacteria and plants via anaplerotic replenishing of C4-dicarboxylic acid into the tricarboxylic acid cycle (TCA). The synthesis of OAA is a key step in the formation of four-carbon compounds, such as malate, fumarate, and succinate. As such, PEPC performs a central role by maintaining the continuity of the TCA carbon fluxes, connecting glycolysis to the TCA.

PEPC orthologs are present in most bacterial lineages (COG2352, http://www.ncbi.nlm.nih.gov/COG/new/release/cow.cgi?cog= COG2352[2]) and are widespread in plants. In contrast, no PEPC homologs have been detected in Archaea so far. However, PEPC activities have been measured[3] and the corresponding enzymes were purified and biochemically characterized in two archaeal hyperthermophiles, *Methanothermus sociabilis*[4] and *Sulfolobus acidocaldarius*[5]. In the latter studies, it has been shown that the archaeal enzyme resembles the bacterial PEPC in quaternary structure, being a tetramer that requires Mg^{2+} for its activity. However, unlike the bacterial enzyme, the archaeal PEPC is considerably smaller in size and lacks some typical regulatory properties[4,5]. Interestingly, a recent comparative genomic study of central carbon metabolic pathways of autotrophic methanogens failed to identify a gene encoding an apparent candidate to connect pyruvate metabolism to the partial reductive TCA cycle in the genome of *Methanopyrus kandleri*[6]. Obviously, this gene remained to be detected.

Here we describe the computational prediction of a novel, divergent archaeal-type PEPC family (atPEPC), which is encoded in most of the sequenced archaeal genomes, including *Methanopyrus kandleri*, and the functional characterization of atPEPC from *Sulfolobus solfataricus* P2. The properties of the recombinantly produced atPEPC closely resembled the characteristics of the PEPC enzyme that was previously purified from the closely related species *Sulfolobus acidocaldarius*.

RESULTS AND DISCUSSION

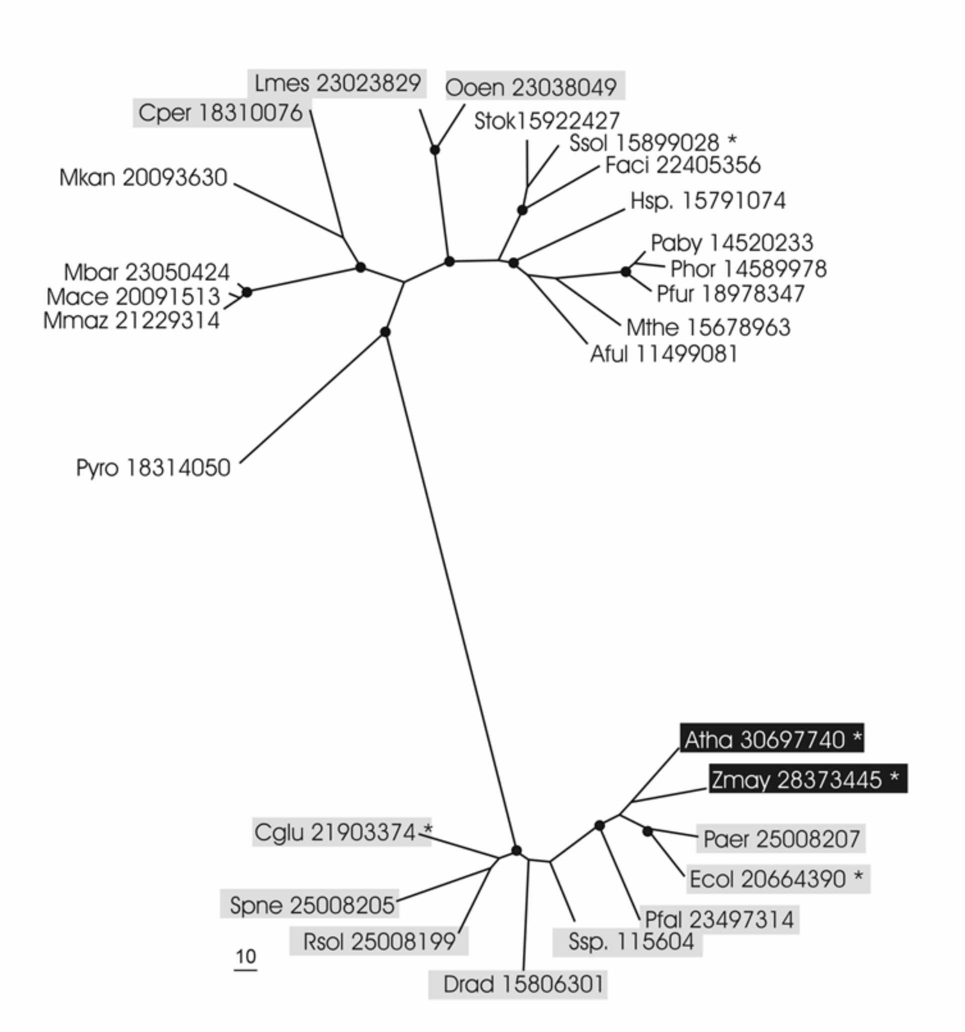


Figure 7.1 Maximum likelihood tree of atPEPC and BE-PEPC sequences. Nodes with bootstrap probability >70% are marked by circles. Archaeal proteins are in plain text, eukaryotic proteins are shaded black, and bacterial proteins are shaded grey. The sequences are denoted by their GI numbers and abbreviated species names. Species abbreviations: Pyro: *Pyrobaculum aerophilum*; Ssol: *Sulfolobus solfataricus*; Stok: *Sulfolobus tokodaii*; Hsp: *Halobacterium* sp.; Mace: *Methanosarcina acetivorans*; Mmaz: *Methanosarcina mazei*; Mbar: *Methanosarcina barkeri*; Mkan: *Methanopyrus kandleri*; Mthe: *Methanothermobacter thermautotrophicus*; Paby: *Pyrococcus abyssi*; Pfur: *Pyrococcus furiosus*; Phor: *Pyrococcus horikoshii*; Aful: *Archaeoglobus fulgidus*; Faci: *Ferroplasma acidarmanus*; Atha: *Arabidopsis thaliana*; Zmay: *Zea mays*; Pfal: *Plasmodium falciparum*; Ooen: *Oenococcus oeni*; Cper: *Clostridium perfringens*; Lmes: *Leuconostoc mesenteroides*; Paer: *Pseudomonas aeruginosa*; Ecol: *Escherichia coli*; Ssp: *Synechococcus* sp.; Drad: *Deinococcus radiodurans*; Rsol: *Ralstonia solanacearum*; Spne: *Streptococcus pneumoniae*; Cglu: *Corynebacterium glutamicum*. An asterisk marks characterized proteins.

Computational identification of the archaeal homologs of PEPC

In the course of a systematic analysis of archaeal central carbon metabolic pathways[7-10], we detected significant similarity between archaeal proteins from COG1892 and the bacterial/eukaryote PEPC family (BE-PEPC; COG2352) using the PSI-BLAST program[11]. For example, a PSI-BLAST search that was seeded with *S. solfataricus* sequence SSO2256 (gi|15899028) from COG1892, retrieved a PEPC-like ortholog from *Thermus* sp. (gi|1581933), at the second iteration with E-value of $1 \cdot 10^{-6}$. A reverse search that is initiated with the PEPC of *E. coli* (gi|20664390) retrieves PF1975 (gi|18978347), the member of COG1892 from *Pyrococcus furiosus*, with $E=4 \cdot 10^{-4}$ at the second iteration, further supporting the connection between these protein families. In addition to species listed in COG1892, we detected orthologous proteins encoded in the genomes of several archaeal species (*P. furiosus, Methanosarcina barkeri* and *M. mazei, Ferroplasma acidarmanus*, *Sulfolobus tokodaii, Picrophilus torridus;* Table I) and three gram-positive Bacteria (*Clostridium perfringens, Oenococcus oeni,* and *Leuconostoc mesenteroides).* Most likely, these bacteria have acquired the archaeal PEPC gene via lateral gene transfer, since these bacterial sequences confidently cluster within the archaeal sequences in a maximum likelihood-based phylogenetic tree (Figure 7.1). This seems to be a case of non-orthologous gene displacement[12] where the archaeal enzyme apparently has displaced the typical bacterial PEPC. It should be noted that proteins belonging to this orthologous group have been hypothesized to be PEPC before[9].

Sequence analysis of atPEPC

The most notable difference between bacterial/eukaryotic PEPC (BE-PEPC) and archaeal-type PEPC (atPEPC) is the apparent molecular weight of the respective proteins. The mass of a typical BE-PEPC ranges from 90 – 110 kDa, whereas the calculated molecular masses of atPEPC subunits are approximately half this size, ranging from 55 – 60 kDa, which is in good agreement with the data for the archaeal PEPCs that have been purified and biochemically characterized (60 ± 5 kDa)[4,5]. Analysis of the multiple alignment of the two families shows that the difference in size between BE-PEPC and atPEPC is mainly due to large insertions in BE-PEPC in the regions surrounding Motif I (Figure 7.2). According to the resolved structures of the *E. coli* and maize PEPCs, these insertions form two 4-helical bundles, which are located at the dimer interface and therefore are thought to play a role in the stabilization of the tetramer[13-15]. As has been observed in the maize and *E. coli* PEPC structures, these inserted regions constitute a binding site of a

Table I: Presence of OAA forming and consuming enzymes encoded by archaeal genomes

	PCK A/B	PYC/OAD	atPEPC	MDH[c]	CTS	CTL[e]
Species	*COG1274/1866*[a]	*COG5016/0439*[b]	*COG1892*	*COG0039/2055*	*COG0372*	*COG2301*
Crenarchaea						
Aeropyrum pernix	-/APE0033	-/-	-	APE0672/-	APE1713	APE0311
Pyrobaculum aerophilum	-/-	-/-	PAE3416	PAE2370/-	PAE3585+ PAE3586[d], PAE1689	-
Sulfolobus solfataricus	SSO2537/-	-/-	SSO2256	SSO2585/-	SSO2589	SSO1254
Sulfolobus tokodaii	ST1058/-	-/-	ST2101	ST1811/-	ST1805, ST0587	-
Euryarchaea						
Archaeoglubus fulgidus	-/-	AF1252/ AF0220	AF1486	AF0855/-	AF1340	-
Halobacterium sp.	-/-	-/-	VNG2259C	VNG2367G/-	VNG2102G	VNG0627G
Methanosarcina acetivorans	-/-	MA0674/ MA0675	MA2690	MA0819/-	MA0249	-
Methanosarcina barkeri	-/-	Meth02001244/ Meth02001245	Meth02002803	Meth02003982/-	Meth02000729	-
Methanosarcina mazei	-/-	MM1827/ MM1828	MM3212	MM1966/-	MM1527	-
Methanococcus jannaschii	-/-	MJ1231/ MJ1229	-	MJ0490/ MJ1425	-	-
Methanopyrus kandleri	-/-	-/-	MK0190	MK1069/ MK0392	-	-
Methanothermobacter thermoautotrophicus	-/-	MTH1107/ MTH1917	MTH943	MTH188/ MTH1205	MTH1726, MTH962	-
Pyrococcus abyssi	PAB1253/-	-/-	PAB2342	-/PAB1791	-	-
Pyrococcus furiosus	PF0289/-	-/-	PF1975	-/-	PF0203	-
Pyrococcus horikoshii	PH0312/-	-/-	PH0016	-/PH1277	-	-
Ferroplasma acidarmanus	Faci02001897/-	-/-	Faci02001756	Faci02000699/-	Faci02000612, Faci02001586	-
Picrophilus torridus	-/-	-/-	PTO0964	PTO0994/-	PTO0889, PTO0169	-
Thermoplasma acidophilum	TA0123/-	-/-	-	TA0952/-	TA0169, TA0819	-
Thermoplasma vulcanium	TVN0200/-	-/-	-	TVN1097/-	TVN0239	-

[a] There are two distinct types of PEP carboxykinases, belonging to different COGs: COG1274 contains ATP dependent PEP carboxykinases, COG1866 contains the GTP dependent PEP carboxykinases.[b] The archaeal pyruvate carboxylase/oxaloacetate decarboxylase consists of a heterodimeric protein[16]: COG5016 contains the enzymatic α-subunit and COG0439 the biotinylated β-subunit. Only if both the enzymatic (COG5016) and the biotinylated subunit (COG0439) are present, the corresponding protein identifiers are indicated. [c] COG0039 encodes the bacterial-type and COG2055 the archaeal-type[17] malate dehydrogenase. [d] PAE3585 and PAE3586 appear to encode a single, full-length citrate synthase gene which is disrupted by a genuine frame shift. [e] COG2301 encodes the β-subunit of the ATP:citrate lyase. The α and γ subunits are not encoded by archaeal genomes. For enzyme abbreviations see figure 7.4. Absence of a gene ortholog is depicted with '–'.

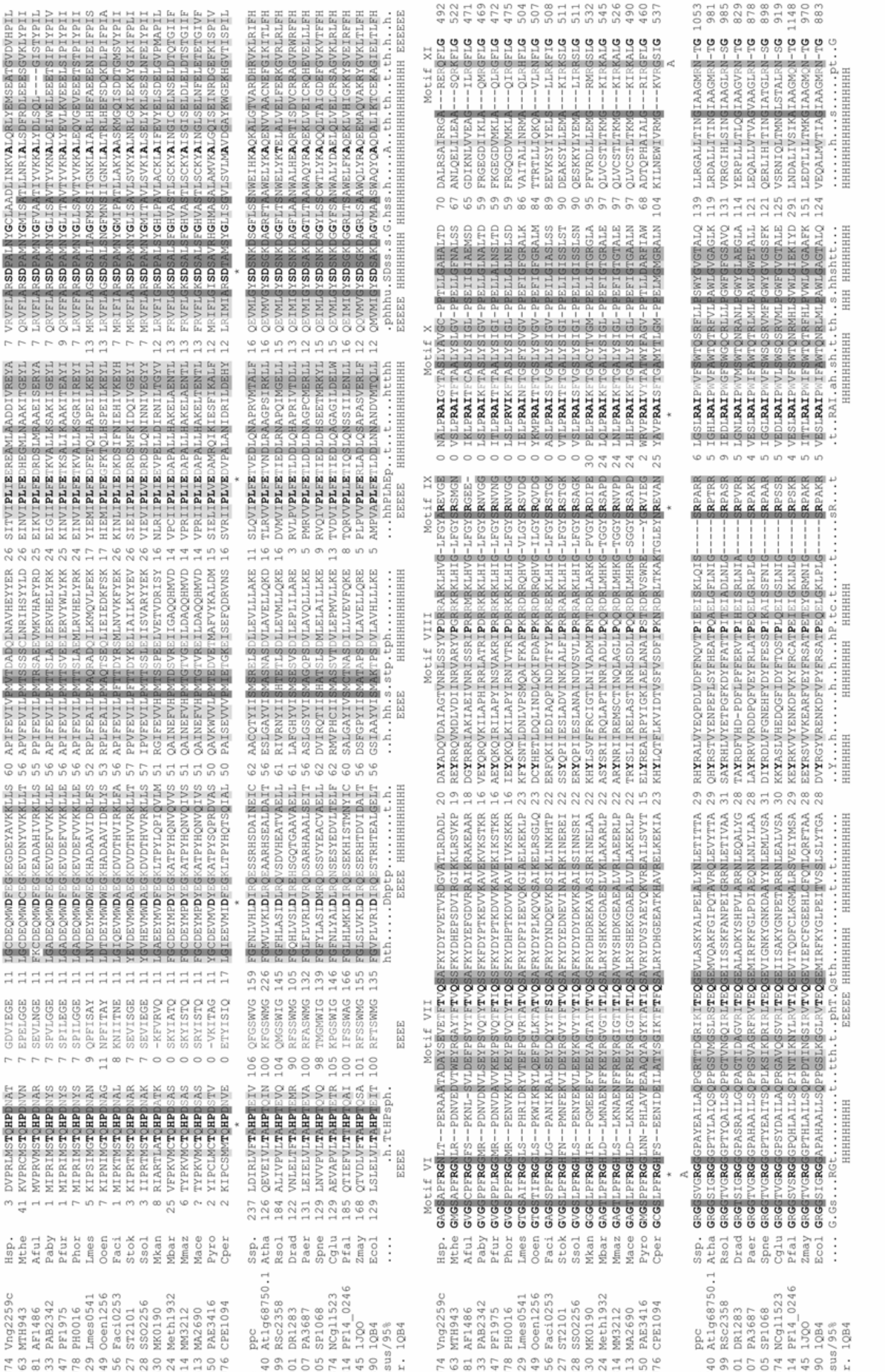

Motif I
Motif II
Motif III
Motif IV
Motif V
Motif VI
Motif VII
Motif VIII
Motif IX
Motif X
Motif XI
consensus/95%
Sec.str. 1QB4

Figure 7.2 (Previous page; NB: For COLOR version of this figure, see APPEDIX I) Multiple alignment of the conserved core of archaeal, bacterial and eukaryote PEPC sequences. The sequences are denoted by Gene Identification (gi) numbers from the GenBank database, species abbreviation (see legend to Figure 7.1) and systematic gene numbers; proteins with available structure denoted by their PDB code. The positions of the first and the last residue of the aligned region in the corresponding protein are indicated for each sequence. The numbers within the alignment represent poorly conserved inserts that are not shown. Amino acids residues that are involved in the formation of the active centre are shown by asterisks; those that have an anticipated involvement in L-aspartate binding are denoted as "A" in the line between atPEPC and BE-PEPC families. Positions with identical amino acids in both families are boldfaced. The coloring is based on the consensus (calculated for all sequences in the alignment) shown underneath the alignment; *h* indicates hydrophobic residues (*ACFILMVWYH*), *t* indicates turn-forming residues (*ASTDNVGPENRK*), *p* indicates charged residues (*STEDKRNQH*), *s* indicates small residues (*AGSVC*), *a* indicates aromatic residues (*FYW*). The secondary structure elements correspond to those experimentally identified for 1QB4[15]. *H* indicates α-helix; *E*, β-strand. Sequence region of the β-strand 2 for 1QB4 and predicted by JPRED program[18] for gi|18310076 (CPE1094) as a query β-strand 2 for atPEPC are shown in blue.

sulphate ion, which was located in the proposed binding site of BE-PEPC allosteric activator, glucose 6-phosphate[13,14]. Indeed, three out of four conserved arginine residues (R183, R184, R231 and R372) of bacterial and eukaryotic PEPCs, which are notably absent in the archaeal sequences, have been shown to be involved in allosteric regulation by glucose-6-P, emphasizing the importance of this region in modulating the enzymatic activity[14]. L-aspartate is a well-known allosteric inhibitor of BE-PEPC, and the residues R587, K773, R832 and Q881 have been identified as its binding pocket[15]. Of the four amino acids responsible for L-aspartate binding, only the catalytically important R587 (motif VI) is conserved in atPEPC. In the *E. coli* PEPC, L-aspartate binding renders the enzyme inactive by dislocating R587. Two other residues, K773 and R832 of PEPC *E. coli,* from are located in the region between motif X and XI and are only conserved throughout the BE-PEPC family. There are no conserved residues in the archaeal family in this region (data not shown).

The rest of the protein is tightly packed around an eight-strand β-barrel and includes the active site and the PEP-binding motif[13,15,19]. In the SCOP database[20], this domain is classified as a TIM barrel fold, superfamily PEP/pyruvate domain, which is also found in pyruvate kinase, pyruvate phosphate dikinase and a few other enzymes that are capable of converting structurally similar substrates (http://scop.mrc-lmb.cam.ac.uk/scop/data/scop.b.d.b.bd. html). Seven of the 8 conserved β-strands of the BE-PEPC family could be confidently identified in atPEPC (see motifs I-VII in Figure 7.2), with the sole exception of strand-2. However, the atPEPC family proteins have a predicted β-strand between motifs I and II that could potentially occupy the position of strand-2 in the atPEPC structure (Figure 7.2). Most of the catalytically important amino acid residues of BE-PEPCs are conserved in atPEPCs, except for the single substitution of a conserved arginine residue that is replaced by a glutamic acid in motif II of the archaeal family (Figure 7.2).

Taken together, the results of these computational analyses suggest that the archaeal homologs of BE-PEPCs have the same enzymatic activity but are unlikely to be subject to the same type of allosteric regulation.

Heterologous production and purification of atPEPC

Extensive effort to functionally produce several atPEPC candidates from different archaeal species using T7 expression systems were unsuccessful due to the formation of inclusion bodies (not shown). Subsequently, an attempt was made to enhance the solubility by producing atPEPC fused to the *E. coli* maltose binding protein (MBP). The ability of MBPs to enhance the solubility of similar 'hard-case' proteins has been well studied[21]. In order to create a MBP fusion protein, ORF SSO2256 from *S.*

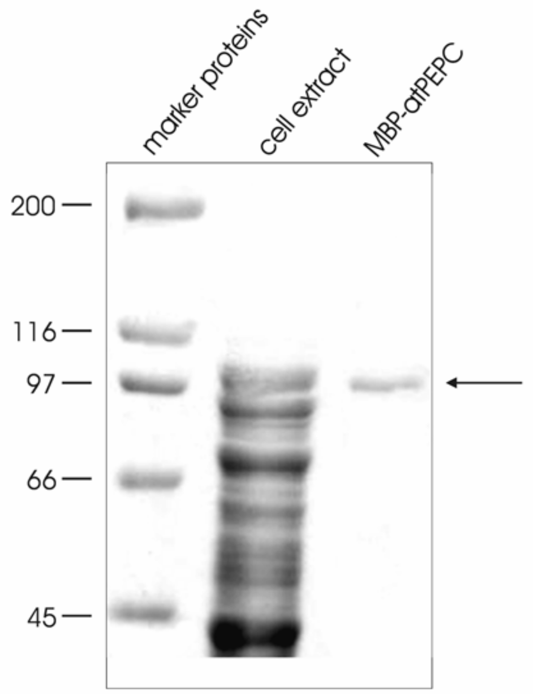


Figure 7.3 SDS-PAGE of purified MBP-atPEPC fusion protein. The SDS-PAGE was performed at a polyacrylamide concentration of 8% and with either 500 μg of cell extract of an expression culture or 10 μg of purified fusion protein, which was obtained using affinity chromatography (see methods). The arrow indicates a single band of approximately 100 kDa, closely corresponding to the calculated molecular weight of the fusion protein (101 kDa).

solfataricus, encoding the atPEPC candidate, was PCR-amplified from genomic DNA from *Sulfolobus solfataricus* P2 and cloned into the *E. coli* MBP-fusion vector pMAL-c2T, resulting in pWUR140. The atPEPC candidate from *S. solfataricus* was functionally expressed as a MBP fusion protein and could be purified from cleared lysates from an expression culture by affinity chromatography using an amylose resin (New England Biolabs, MA, USA) (Figure 3). After elution from amylose resin with elution buffer, the MBP-atPEPC fusion protein eluted as a single band with an estimated molecular mass of approximately 100 kDa according to SDS-PAGE analysis. Several attempts to separate MBP from the atPEPC enzyme by proteolytic cleavage, using the thrombin site that is present in the linker region in between the fused proteins, were unsuccessful. Exposure of the MBP-atPEPC fusion protein to varying amounts of thrombin resulted in a-specific cleavage of the fusion protein and inactivation of the PEPC activity (data not shown). Due to these problems, we decided to perform a brief biochemical characterization of atPEPC using the purified MBP-atPEPC fusion protein.

Physical and biochemical characterization of atPEPC

The atPEPC from *S. solfataricus* displayed the highest activity at around 85°C (Table II), which is close to the optimal temperature reported for the purified PEPC from *S. acidocaldarius* (90°C)[5]. In addition, the enzyme activity was completely dependent on the presence of Mg^{2+}. The enzyme activity could not be reconstituted by addition of Mn^{2+} instead of Mg^{2+} (Table II). Morover *S. solfataricus* atPEPC displayed Michaelis–Menten kinetics with a $K_{m[PEP]}$ and specific activity of 0.09 mM and 2.1 U/mg, respectively under the standard assay conditions (Table II). These observations are in good agreement with the data reported for the purified *S. acidocaldarius* enzyme[5]. Given the reported lack of allosteric regulation of the previously characterized archaeal PEPCs by the known effectors of BE-PEPC, we analyzed the effect of these effectors on the recombinant *S. solfataricus* atPEPC. Allosteric activators of *E. coli* PEPC (glucose-6-P, fructose 1,6-bisphosphate, acetyl-CoA) did not affect the activity of atPEPC (Table II). This lack of allosteric activation could be explained by the absence, in atPEPC sequences, of counterparts of the regions of BE-PEPC that are involved in allosteric activation (see above). However, addition of L-aspartate and L-malate to a final concentration of 5 mM to the standard enzyme assay resulted in inhibition of PEPC activity of 50% and 20%, respectively, which roughly corresponds to the characteristics of the *S. acidocaldarius* enzyme[5]. The sensitivity of both *Sulfolobus* PEPC

Table II: Comparison of characterized archaeal type and BE-PEPCs

	Archaea			**Bacteria**	**Eukarya**
Parameter	*S. solfataricus*	*S. acidocaldarius*	*M. sociabilis*	*E. coli*	*Z. mays*
Molecular weight(kDa) /number of subunits	- [a]	260/4	240/4	360/4	400/4
MW per monomer (kDa)	58[b]	60	60	90	100
T_{opt}	85	90	85	35-38	40
K_m *(PEP)*	0.09	0.2	1.3	20	0.1
V^{max} *(U/mg)*	2.1	2.77	2.70	30	20.8
Optimum pH	8.0	8.0	8.5	7.5	7.5 - 8.0
Allosteric inhibitors	L-aspartate, L-malate	L-aspartate, L-malate	-	L-aspartate, L-malate, citrate, succinate, fumarate	L-aspartate, L-malate, citrate, pyruvate, oxaloacetate
Allosteric activators	-	-	-	Ac-CoA, FBP, fatty acids, GTP	G6P, G1P, glycine
Metals	Mg^{2+}	Mg^{2+}	Mg^{2+}, Mn^{2+}	Mg^{2+}	Mg^{2+}
Reference	This study	5	4	22	23

[a] The molecular weight and multimeric conformation of the native MBP-PEPC could not be determined accurately, since the fusion protein was found to be associated in a large complex (Ettema, T. J. G., unpublished observation). **[b]** Calculated molecular weight. Abbreviations: Ac-CoA: acetyl coenzyme A; FBP: fructose 1,6-bisphosphate; G6P: glucose 6-phosphate; G1P: glucose 1-phosphate.

activities to L-aspartate is surprising given the absence of known allosteric domains. Conceivably, atPEPCs evolved a distinct mechanism of L-aspartate binding and regulation, which is compatible with the low degree of conservation of the predicted L-aspartate-contacting residue in motif XI in the atPEPC family (Figure 7.2). Purified PEPC from the archaeal methanogen *Methanothermus sociabilis* has been reported to be insensitive to L-aspartate-mediated repression, as well as to all other known PEPC effectors, including L-malate[4]. The sequence of this particular atPEPC, which is currently unavailable, might provide clues as to the mechanisms of allosteric regulation in this archaeal protein family.

Functional implications of atPEPC in archaeal central carbohydrate pathways

The detection of a gene encoding PEPC activity sheds light on the organization of the central carbon metabolism in Archaea, especially in autotrophic species. Most sequenced archaeal genomes encode a partial TCA cycle[9] that is used as a source of intermediates (e.g. OAA, succinate and α-ketoglutarate) for the biogenesis of amino acids and vitamins. For some methanogenic species, e.g. the *Methanococcales* and *Methanobacteriales*, a partial TCA has been found to operate in a reductive direction from OAA to α-ketoglutarate[24]. In contrast, methanogens from the order of *Methanosarcinales* make use of a partial TCA that operates in the oxidative direction from OAA to α-ketoglutarate[24]. Therefore, it is obvious that OAA biosynthetic pathways play a central role in methanogenic archaea. The identification of the archaeal PEPC gene solves the problem of OAA formation for *Methanopyrus kandleri*[6]. This finding also suggests that in methanogenic archaea and *A. fulgidus* atPEPC is an important link between gluconeogenisis and amino acid biosynthesis. This hypothesis is indirectly supported by the fact that citrate synthase is not encoded in the *M. kandleri* and *M. jannaschii* genomes (Table I). It should be noted however, that, apart from the atPEPC gene, three archaeal methanogens and *A. fulgidus* additionally have genes encoding a heterodimeric pyruvate carboxylase[16,25] (COG5016/439), that is anticipated to be involved in oxaloacetate formation (Table I and Figure 7.4a).

Whereas atPEPC is probably involved in OAA formation as part of a reductive TCA cycle in autotrophic Archaea, their most likely role in heterotrophic Archaea is solely anaplerotic. In the pyrococcal genomes, the atPEPC gene forms a putative operon, which additionally contains genes involved in L-aspartate and NAD biosynthesis in all three (Figure 7.4b). Aspartate is an important precursor in the biosynthetic routes of several other amino acids, such as asparagine, methionine, leucine, and isoleucine (Figure 7.4a), which makes aspartate a well-suited signal molecule for the demand for these amino acids in the cell. Not surprisingly, most PEPCs characterized to date, including the atPEPCs from heterotrophic Archaea, are subject to aspartate-mediated allosteric repression. An anaplerotic role for atPEPC in heterotrophic Archaea is suggested also by the recent whole-genome microarray analysis of *P. furiosus*[26]. The PEPC gene in this organism (PF1975) was 20-fold up-regulated in a maltose-grown culture compared to a culture that was grown on peptides.

The identification, characterization and comparison of enzymes from the archaeal variants of the Embden-Meyerhof, Entner-Doudoroff and TCA pathways have revealed several instances of non-orthologous gene displacement as well as some unique genes (for reviews see references 10 and 9). The identified archaeal-type PEPC, with its distinct properties, constitutes yet another example of the versatility of the enzymes of the central carbon metabolic pathways in the archaeal domain.

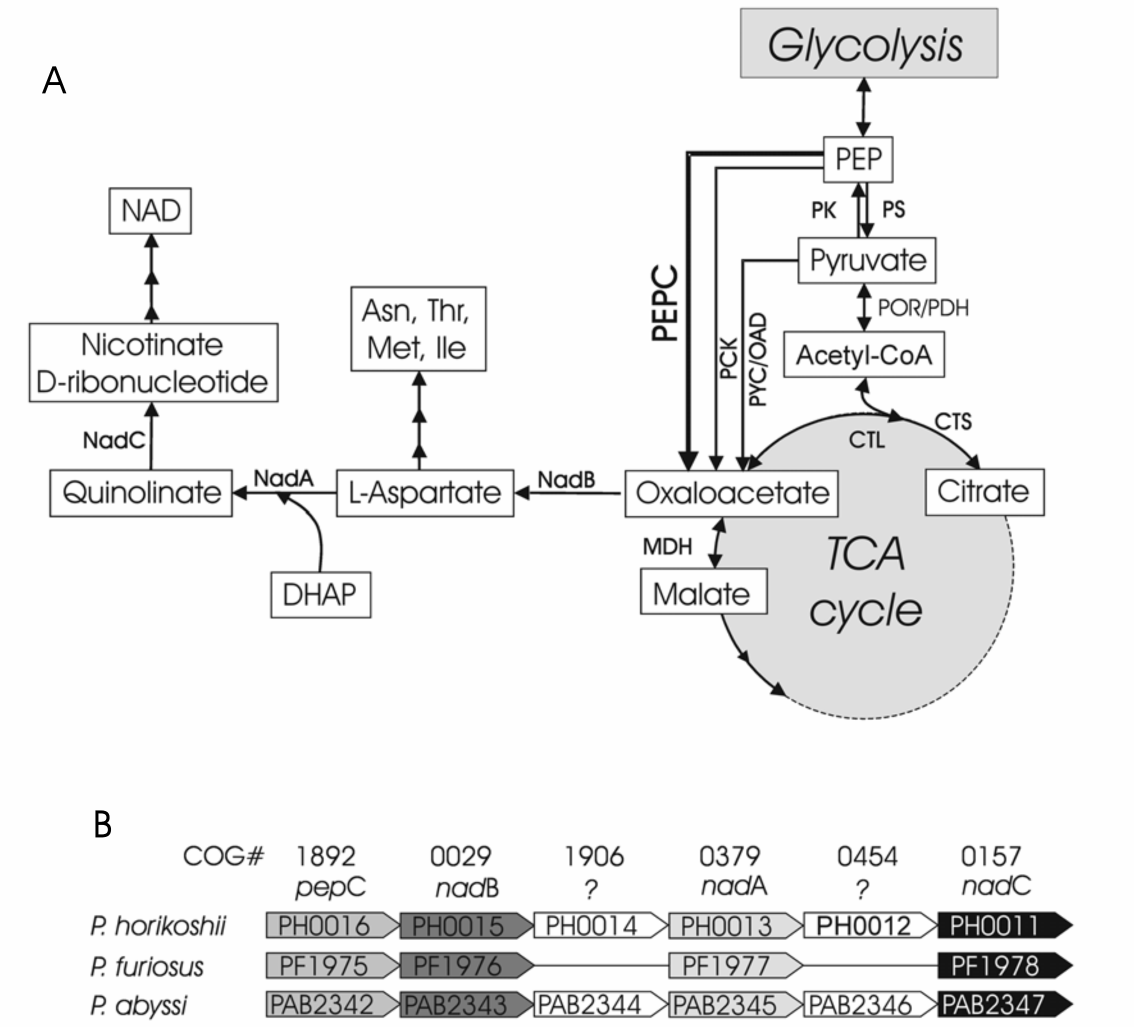


Figure 7.4 (a) Overview of oxaloacetate forming and consuming pathways in Archaea. Abbreviations: PEPC: PEP carboxylase; PYC: pyruvate carboxylase; OAD: oxaloacetate carboxylase; PCK: PEP carboxykinase (ATP/GTP dependent); PK: pyruvate kinase; PS: PEP synthase; CTS: citrate synthase; CTL: ATP:citrate lyase; POR: pyruvate oxido-reductase; PDH: pyruvate dehydrogenase; NadA: quinolinate synthase; NadB: L-aspartate oxidase; NadC: nicotinate-nucleotide pyrophosphorylase (carboxylating); DHAP: dihydroxy-acetone-phosphate. **(b)** Conserved local gene context of atPEPC and aspartate/NAD biosynthesis genes pathway in *Pyrococcus*. Orthologous genes are shaded in the same grey-scale and corresponding COGs are depicted; genes encoding proteins with an unknown function are boxed white. Genes are not drawn to scale.

EXPERIMENTAL PROCEDURES

Database searches

In order to detect sequences that are distantly related to the bacterial and eukaryote PEPC sequences in the archaeal domain, the Non-Redundant database (NRDB) of protein sequences (National Center for Biotechnology Information, NIH, Bethesda; http://www.ncbi.nih.gov/BLAST) was searched using the PSI-BLAST program, using the BLOSUM62 matrix and a cut-off of E=0.01 for inclusion of sequences in position-specific scoring matrices[11,27].

Multiple alignments and phylogenetic tree construction

Multiple alignments of protein sequences were constructed using the T-Coffee program[28], followed by manual adjustment for conserved motifs based on the PSI-BLAST results. Protein secondary structure was predicted using the JPRED program[18]. The same alignment was used for phylogenetic tree reconstruction as follows. Evolutionary distances for the trees reconstruction were calculated from multiple sequence alignments using the Dayhoff PAM model as implemented in the PROTDIST program of the PHYLIP package[29]. Distance trees were constructed using the least-square method[30] as implemented in the FITCH program of PHYLIP[29]. Maximum likelihood trees were constructed by using the ProtML program of the MOLPHY package, with the JTT-F model of amino acid substitutions[31,32], to optimize the least-square trees with local rearrangements. Bootstrap analysis was performed for the maximum likelihood tree as implemented in MOLPHY using the Resampling of Estimated Log-Likelihoods (RELL) method[32,33].

Organisms and growth conditions

Escherichia coli JM109 was used both for cloning and subsequent functional expression of target genes and was cultivated in Luria Bertani medium containing ampicillin (100 μg/ml) in a rotary shaker at 37 °C.

Cloning and functional expression of atPEPC from *Sulfolobus solfataricus*

The method described above identified an atPEPC candidate in the genome of *S. solfataricus*, open reading frame SSO2256 (gi|15899028, SSO-

PEPC), which subsequently was chosen for recombinant expression in *Escherichia coli*. In order to create a MBP-atPEPC fusion protein, ORF SSO2256 was PCR-amplified using the primer pair BG1586 (5'-GCGCG**GAATTC**ATGAGAATCATACCACGCACTATGTC-3') and BG1587 (5'-GCGCG**GTCGAC**TCATCCCAAGGATCTTCTAATTAATGC-3'), with the introduced *Eco*RI and *Sal*I restriction sites in bold. The resulting PCR product was digested and cloned into the MBP-fusion vector pMAL-c2T[34], which was kindly provided by P. Riggs (New England Biolabs, MA, USA). The resulting plasmid, pWUR140, was used for recombinant expression in *E. coli* JM109 using standard procedures. A 1 litre culture of *E. coli* JM109 harboring the pWUR140 plasmid was grown and expression was at induced $OD_{595} = 0.5$ by addition of IPTG (final concentration 0.4 mM). After an additional incubation at 37 °C for 4 hours, allowing expression of heterologous protein, cells were harvested by centrifugation (2,200 x g for 10 min) and resuspended in 10 ml of column buffer (50 mM Tris-HCl, pH 8.0, 200 mM NaCl, 1mM EDTA). Cells were then lyzed by sonication (Branson), during which the sample was cooled on ice/ethanol slurry. Cell debris was removed by centrifugation (10,000 x g for 20 min). The resulting supernatant was used for purification of the recombinant atPEPC.

Purification of recombinant atPEPC

Cell-free extract from a 1-litre expression culture was applied to 1 ml (bed volume) amylose resin (New England Biolabs), which was pre-equilibrated with 5 column volumes of column buffer. Subsequently, the amylose resin was washed with 10 column volumes of column buffer to wash out unbound proteins. Finally, highly purified, active MBP-atPEPC fusion protein could be eluted from the amylose resin by adding 1 column volume of elution buffer (column buffer containing 10 mM of maltose). Routinely, a 1-litre expression culture yielded approximately 0.5 mg of purified MBP-atPEPC.

Protein concentration and purity

Protein concentrations were determined with Coomassie Brilliant Blue G-250 as described before[35] using bovine serum albumin as a standard. The purity of the fusion protein was checked by SDS-PAGE. Protein samples for SDS-PAGE were heated for 5 min at 100 °C in an equal volume of sample loading buffer (0.1 M citrate-phosphate buffer, 5% SDS, 0.9% 2-mercaptoethanol, 20% glycerol, pH 6.8).

Standard PEPC enzyme assay

PEPC activity was routinely measured in a coupled enzyme assay using the thermostable malate dehydrogenase (MDH) from *Thermus flavus* (Sigma) by spectophotometrically monitoring the NADH oxidation at 340 nm at a Hitachi 2010 Spectophotometer at 80 °C. All enzyme assays, unless stated otherwise, were started by adding the substrate (PEP) to a final concentration of 5 mM to the preheated reaction mix, containing purified recombinant *S. solfataricus* MBP-atPEPC fusion protein (20 µg/ml), 50 mM Tris-HCl pH 8.0, 10 mM Na_2CO_3, 2 mM $MgSO_4$, 0.15 mM NADH and 2 U of *T. flavus* MDH in a final volume of 1 ml. One unit of PEPC activity was defined as the amount of enzyme that is needed for the conversion of 1 µmol of NADH per minute.

Effects of metal ion concentration, potential allosteric effectors of atPEPC activity

Known allosteric activators (glucose 6-phosphate, fructose 1,6-bisphosphate) and inhibitors (L-aspartate, L-malate, acetyl-CoA) of BE-PEPC were tested on the activity of the *S. solfataricus* atPEPC by adding varying amounts (0–5 mM) to the standard enzyme assays at 80 °C. Metal ion dependence of atPEPC activity was determined by adding different amounts (0-10 mM) of either $MgSO_4$ or $MnSO_4$ to the standard enzyme assay mixture lacking divalent metal ions.

ACKNOWLEDGEMENT

We thank Paul Riggs (New England Biolabs, Boston, USA) for kindly providing plasmid pMAL-c2T.

NOTE ADDED IN PROOF

While the present study was in review, a report by Patel *et al.*[36] was published that describes the purification and characterization of an archaeal PEPC of *Methanothermobacter thermoautotrophicus*. By determination of the N-terminal sequence, these authors identified the archaeal PEPC family that is also reported here. It should be noted that both the strategy of Patel et al. (a reverse genetics approach) and the strategy used here (a function prediction/verification via bioinformatics approach) resulted in the identification of the same archaeal PEPC family.

REFERENCES

1 Chollet, R., Vidal, J. and O'Leary, M.H. (1996) Phospho*enol*pyruvate carboxylase: A ubiquitous, highly regulated enzyme in plants, *Annu Rev Plant Physiol Plant Mol Biol* **47**, 273-298

2 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., Rao, B.S., Smirnov, S., Sverdlov, A.V., Vasudevan, S., Wolf, Y.I., Yin, J.J. and Natale, D.A. (2003) The COG database: an updated version includes eukaryotes, *BMC Bioinformatics* **4**, 41

3 Zeikus, J.G., Fuchs, G., Kenealy, W. and Thauer, R.K. (1977) Oxidoreductases involved in cell carbon synthesis of *Methanobacterium thermoautotrophicum*, *J Bacteriol* **132**, 604-13

4 Sako, Y., Takai, K., Uchida, A. and Ishida, Y. (1996) Purification and characterization of phospho*enol*pyruvate carboxylase from the hyperthermophilic archaeon *Methanothermus sociabilis*, *FEBS Lett* **392**, 148-52

5 Sako, Y., Takai, K., Nishizaka, T. and Ishida, Y. (1997) Biochemical relationship of phospho*enol*pyruvate carboxylases (PEPCs) from thermophilic archaea, *FEMS Microbiology Letters* **153**, 159-165

6 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., Tatusov, R.L., Wolf, Y.I., Stetter, K.O., Malykh, A.G., Koonin, E.V. and Kozyavkin, S.A. (2002) The complete genome of hyperthermophile *Methanopyrus kandleri* AV19 and monophyly of archaeal methanogens, *Proc Natl Acad Sci U S A* **99**, 4644-9

7 Makarova, K.S. and Koonin, E.V. (2003) Comparative genomics of archaea: how much have we learned in six years, and what's next?, *Genome Biol* **4**, 115

8 Makarova, K.S. and Koonin, E.V. (2003) Filling a gap in the central metabolism of archaea: prediction of a novel aconitase by comparative-genomic analysis, *FEMS Microbiol Lett* **227**, 17-23

9 Huynen, M.A., Dandekar, T. and Bork, P. (1999) Variation and evolution of the citric-acid cycle: a genomic perspective, *Trends Microbiol* **7**, 281-91

10 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**, 231-46

11 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**, 3389-402

12 Koonin, E.V., Mushegian, A.R. and Bork, P. (1996) Non-orthologous gene displacement, *Trends Genet* **12**, 334-6

13 Matsumura, H., Xie, Y., Shirakata, S., Inoue, T., Yoshinaga, T., Ueno, Y., Izui, K. and Kai, Y. (2002) Crystal structures of C4 form maize and quaternary complex of E. coli phospho*enol*pyruvate carboxylases, *Structure (Camb)* **10**, 1721-30

14 Kai, Y., Matsumura, H. and Izui, K. (2003) Phospho*enol*pyruvate carboxylase: three-dimensional structure and molecular mechanisms, *Arch Biochem Biophys* **414**, 170-9

15 Kai, Y., Matsumura, H., Inoue, T., Terada, K., Nagara, Y., Yoshinaga, T., Kihara, A., Tsumura, K. and Izui, K. (1999) Three-dimensional structure of phospho*enol*pyruvate carboxylase: a proposed mechanism for allosteric inhibition, *Proc Natl Acad Sci U S A* **96**, 823-8

16 Mukhopadhyay, B., Patel, V.J. and Wolfe, R.S. (2000) A stable archaeal pyruvate carboxylase from the hyperthermophile *Methanococcus jannaschii*, *Arch Microbiol* **174**, 406-14

17 Honka, E., Fabry, S., Niermann, T., Palm, P. and Hensel, R. (1990) Properties and primary structure of the L-malate dehydrogenase from the extremely thermophilic archaebacterium *Methanothermus fervidus*, *Eur J Biochem* **188**, 623-32

18 Cuff, J.A., Clamp, M.E., Siddiqui, A.S., Finlay, M. and Barton, G.J. (1998) JPred: a consensus secondary structure prediction server, *Bioinformatics* **14**, 892-3

19 Matsumura, H., Terada, M., Shirakata, S., Inoue, T., Yoshinaga, T., Izui, K. and Kai, Y. (1999) Plausible phospho*enol*pyruvate binding site revealed by 2.6 A structure of Mn2+-bound phospho*enol*pyruvate carboxylase from *Escherichia coli*, *FEBS Lett* **458**, 93-6

20 Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol* **247**, 536-40

21 Fox, J.D., Routzahn, K.M., Bucher, M.H. and Waugh, D.S. (2003) Maltodextrin-binding proteins from diverse bacteria and archaea are potent solubility enhancers, *FEBS Lett* **537**, 53-7

22 Teraoka, H., Izui, K. and Katsuki, H. (1972) Phospho*enol*pyruvate carboxylase of *Escherichia coli*: alteration of allosteric properties by photooxidation, *Arch Biochem Biophys* **152**, 821-7

23 Ting, I.P. and Osmond, C.B. (1973) Multiple forms of plant phospho*enol*pyruvate carboxylase associated with different metabolic pathways, *Plant. Physiol.* **51**, 448-453

24 Simpson, P.G. and Withman, W.B. (1993) in *Methanogenesis: ecology, physiology, biochemistry and genetics.* (Ferry, J.G., ed.), pp. 445-472, Chapman and Hall

25 Mukhopadhyay, B., Purwantini, E., Kreder, C.L. and Wolfe, R.S. (2001) Oxaloacetate synthesis in the methanarchaeon *Methanosarcina barkeri*: pyruvate carboxylase genes and a putative *Escherichia coli*-type bifunctional biotin protein ligase gene (*bpl*/*bir*A) exhibit a unique organization, *J Bacteriol* **183**, 3804-10

26 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**, 3935-47

27 Aravind, L. and Koonin, E.V. (1999) Gleaning non-trivial structural, functional and evolutionary information about proteins by iterative database searches, *J Mol Biol* **287**, 1023-40

28 Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment, *J Mol Biol* **302**, 205-17

29 Felsenstein, J. (1996) Inferring phylogenies from protein sequences by parsimony, distance, and likelihood methods, *Methods Enzymol* **266**, 418-27

30 Fitch, W.M. and Margoliash, E. (1967) Construction of phylogenetic trees, *Science* **155**, 279-84

31 Jones, D.T., Taylor, W.R. and Thornton, J.M. (1992) The rapid generation of mutation data matrices from protein sequences, *Comput Appl Biosci* **8**, 275-82

32 Adachi, J. and Hasegawa, M. (1992) in *Computer Science Monographs* (Vol. 27) (Mathematics, I.o.S., ed.)

33 Kishino, H., Miyata, T. and Hasegawa, M. (1990) Maximum likelihood inference of protein phylogeny and the origin of chloroplasts, *J. Mol. Evol.* **31**, 151-160

34 Davis, A.J., Perugini, M.A., Smith, B.J., Stewart, J.D., Ilg, T., Hodder, A.N. and Handman, E. (2004) Properties of GDP-mannose pyrophosphorylase, a critical enzyme and drug target in *Leishmania mexicana*, *J Biol Chem* **279**, 12462-8

35 Bradford, M.M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding, *Anal Biochem* **72**, 248-54

36 Patel, H.M., Kraszewski, J.L. and Mukhopadhyay, B. (2004) The phospho*enol*pyruvate carboxylase from *Methanothermobacter thermautotrophicus* has a novel structure, *J Bacteriol* **186**, 5129-37

Chapter 8

The semi-phosphorylative Entner-Doudoroff pathway in hyperthermophilic archaea — a re-evaluation

Hatim Ahmed*
Thijs J. G. Ettema*
Britta Tjaden
Ans C. M. Geerling
John van der Oost
Bettina Siebers

Biochemical studies have suggested that in hyperthermophilic archaea the metabolic conversion of glucose via the Entner-Doudoroff (ED) pathway generally proceeds via a non-phosphorylative variant. Previously, a key enzyme of the non-phosphorylating ED pathway of *Sulfolobus solfataricus*, 2-keto-3-deoxy-gluconate (KDG) aldolase, has been cloned and characterized. In the present study, a comparative genomics analysis is described that reveals conserved ED gene clusters in both *Thermoproteus tenax* and *S. solfataricus*. The corresponding ED proteins from both archaea have been expressed in Escherichia coli, and their specificity has been identified revealing: (i) a novel type gluconate dehydratase (*gad* gene), (ii) a bi-functional 2-keto-3-deoxy-(6-phospho)-gluconate aldolase (*kdg*A gene), (iii) a 2-keto-3-deoxygluconate kinase (*kdg*K gene), and in *S. solfataricus* (iv) a non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN, *gap*N gene). Extensive *in vivo* and *in vitro* enzymatic analyses indicate the operation of both the semi-phosphorylative and the non-phosphorylative ED pathway in *T. tenax* and *S. solfataricus*. The existence of this branched ED pathway is yet another example of the versatility and flexibility of the central carbohydrate metabolic pathways in the archaeal domain.

Biochemical Journal 390(2), 529-540 (2005)

* Both authors contributed equally

INTRODUCTION

Comparative studies of the glycolysis in thermophilic and hyperthermophilic Archaea have revealed a large number of variations of the classical bacterial and eukaryal routes: the Entner-Doudoroff (ED) pathway and the Embden-Meyerhof-Parnas (EMP) pathway[1-4]. Whereas the ED-like pathway seems to be restricted to the aerobic archaea (e.g. *S. solfataricus*[5] and *Thermoplasma acidophilum*[6]), the archaeal modified EMP pathways are found in most anaerobic archaea (e.g. *Pyrococcus furiosus*, *Thermococcus* sp., *Desulfurococcus amylolyticus* and *Archaeoglobus fulgidus*)[2-4]. The presence of both modified pathways has so far only been demonstrated in the anaerobe *T. tenax*[2,7-10].

The classical ED pathway[11] involves (i) the initial phosphorylation of glucose to glucose 6-phosphate either by a glucokinase or by the action of a phosphoenolpyruvate-dependent phosphotransferase system (PTS), (ii) the oxidation to 6-phosphogluconate by glucose-6-phosphate dehydrogenase and phosphogluconolactonase, (iii) the dehydration to 2-keto-3-deoxy-6-phophogluconate (KDPG) by 6-phosphogluconate dehydratase, and (iv) the cleavage of the characteristic KDPG intermediate by KDPG aldolase yielding glyceraldehyde 3-phosphate (GAP) and pyruvate. GAP is further metabolized via the lower, common shunt of the EMP pathway yielding a second molecule of pyruvate (Figure 8.1).

Whereas the classical pathway seems to be restricted to Bacteria, modifications have been identified in all three domains of life: the Eukarya, Bacteria and Archaea[12]. One of the modified versions of the ED pathway that is generally referred to as the semi-phosphorylative ED pathway[13], concerns (i) the oxidation of glucose to gluconate via glucose dehydrogenase, (ii) the conversion of gluconate by a specific gluconate dehydratase to 2-keto-3-deoxygluconate (KDG), (iii) the subsequent phosphorylation by KDG kinase to form 2-keto-3-deoxy-6-phosphogluconate (KDPG), and (iv) the cleavage by KDPG aldolase (Figure 8.1). The semi-phosphorylative ED pathway has been shown to operate in several species of *Clostridium*[14], as well as the halophilic archaea *Halobacterium saccharovorum* and *H. halobium*[15]. Another variant pathway, the so-called non-phosphorylative ED pathway, has been reported for the hyperthermophilic archaea *S. solfataricus*[5], *S. acidocaldaricus*[8], *T. tenax*[2,7-10], the thermophilic archaeon *T. acidophilum*[6] and several species of the fungal genus *Aspergillus*[16]. In contrast to the semi-phosphorylative ED modification, KDG (rather than KDPG) has been reported to be subjected to aldol-cleavage by the KDG aldolase, forming pyruvate and glyceraldehyde. Glyceraldehyde is further oxidized to form glycerate, either by an NAD(P)$^+$-

dependent glyceraldehyde dehydrogenase[6] or by a ferredoxin-dependent glyceraldehyde oxidoreductase[7,17-19]; glycerate is phosphorylated to 2-phosphoglycerate by glycerate kinase[6]. 2-Phosphoglycerate enters the lower shunt of the EMP pathway and forms a second molecule of pyruvate via the enolase and pyruvate kinase reaction (Figure 8.1).

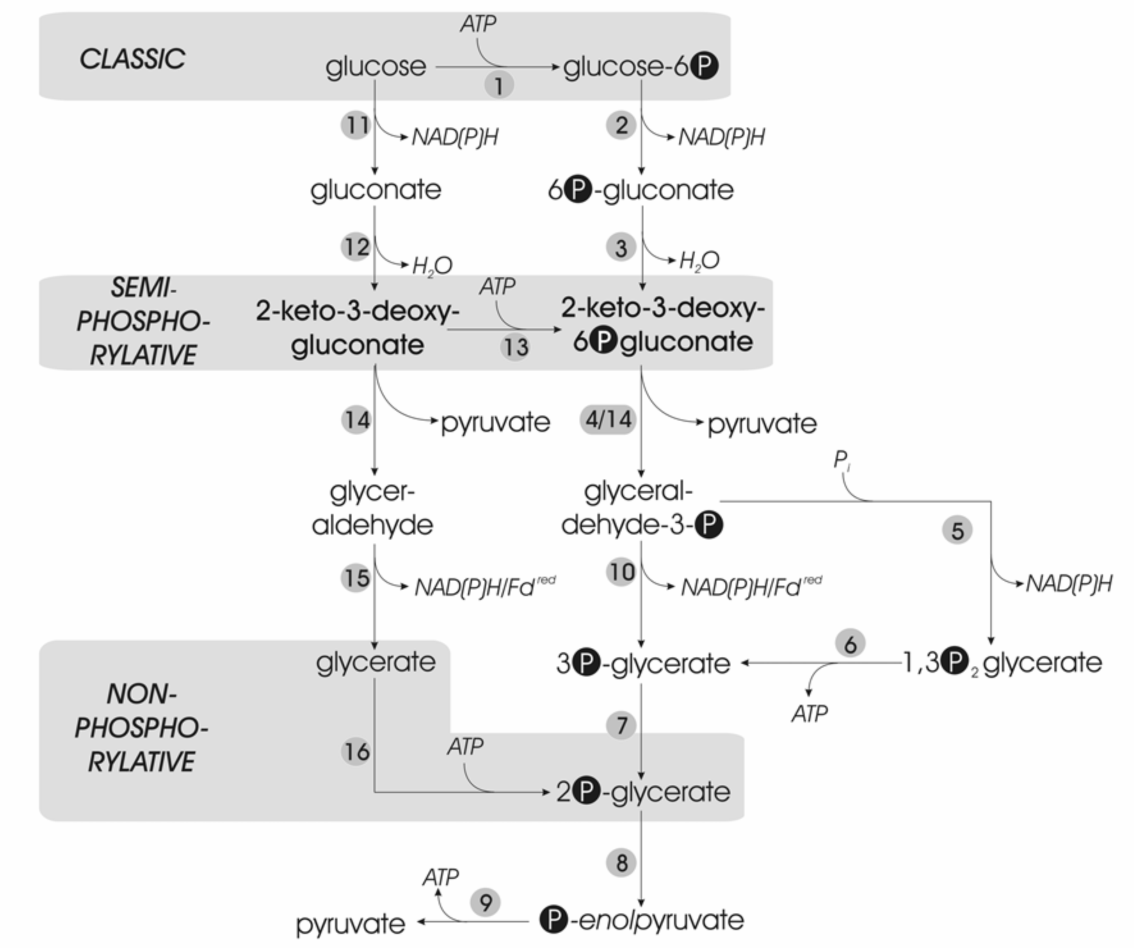


Figure 8.1 Overview of the classical and modifications of the ED pathway, each with the characteristic phosphorylation level indicated. Non-phosphorylated intermediates are depicted on the *left* and phosphorylated intermediates on the *right*. The key phosphorylation reactions for the different ED versions are highlighted in grey boxes (glucokinase/hexokinase for the classical ED, KDG kinase for the semi-phosphorylative ED and glycerate kinase for the non-phosphorylative ED). Key to enzymes: 1, glucokinase/hexokinase; 2, glucose-6-phosphate dehydrogenase; 3, 6-phosphogluconate dehydratase; 4, KDPG aldolase; 5, GAPDH; 6, 3-phosphoglycerate kinase; 7, phosphoglycerate mutase; 8, enolase; 9, pyruvate kinase; 10, GAPN/GAP oxidoreductase; 11, GDH; 12, GAD; 13, KDG kinase; 14, KD(P)G aldolase; 15, aldehyde dehydrogenase/aldehyde oxidoreductase; and 16, glycerate kinase.

In the present study, the ED pathways of two hyperthermophilic Crenarchaea have been re-evaluated: *S. solfataricus* and *T. tenax*.

S. solfataricus grows optimally at 80-85°C and pH 2-4. Aerobic heterotrophic growth is reported on several carbon sources such as starch, glucose, arabinose, fructose and peptide-containing substrates like peptone, tryptone and yeast extract[20]. The non-phosphorylative ED pathway was proposed as pathway for glucose catabolism on the basis of ^{14}C-labelling studies and identification of the characteristic intermediates (KDG and GA)[5], as well as characterization of key enzyme activities[5,21,22]. The glucose dehydrogenase and KDG aldolase of *S. solfataricus* have been studied in detail, and more recent studies[22] indicate that this pathway is promiscuous and represents an equivalent route for glucose and galactose catabolism in this organism. In addition to a glucose dehydrogenase that exhibits high activity with glucose and galactose, the KDG aldolase was shown to lack facial selectivity in catalyzing the cleavage of KDG as well as 2-keto-3-deoxygalactonate (KDGal), both yielding glyceraldehyde and pyruvate. When the present study was submitted, the archaeal gluconate dehydratase had not yet been identified (however, see Discussion).

T. tenax is a sulfur-dependent anaerobe that grows optimally around 90°C, pH 5[23] and was shown to grow both chemolithoautotrophically (CO_2, H_2) and chemoorganoheterotrophically on different carbon sources (e.g. glucose, starch). *T. tenax* uses two different pathways for glucose catabolism, the modified EMP and the non-phosphorylative ED pathway, as deduced from detected enzyme activities in crude extracts, and from the identification of characteristic intermediates in ^{14}C labelling experiments and *in vivo* ^{13}C NMR studies[2,7-10,24]. However, the reconstruction of the central carbohydrate metabolism by the use of genomic and genetic data revealed the presence of the semi-phosphorylative ED pathway in *T. tenax*[2].

In the current view about the ED pathway in Archaea, it is assumed that a semi-phosphorylative version is operative in Haloarchaea, whereas a non-phosphorylative version is present in hyperthermophilic and thermophilic archaea. The aforementioned biochemical data on *S. solfataricus* and *T. tenax* do not disagree with this assumption. However, in our ongoing attempts to reconstruct the archaeal central carbohydrate metabolizing pathways, a comparative genomics approach has revealed ED gene clusters that are conserved in *T. tenax*[2], *S. solfataricus*, *S. tokodaii* and *Halobacterium* sp. NRC. We present here a qualitative analysis of the corresponding gene products of *T. tenax* and *S. solfataricus*. A detailed biochemical analysis of the individual enzymes is ongoing (H. Ahmed, T.J.G. Ettema, J. van der Oost and B.

Sierbers, unpublished work). The present study reports new insights on the operation of the modified ED pathways in hyperthermophilic archaea.

RESULTS AND DISCUSSION

On the basis of biochemical studies using crude extracts as well as characterization of key enzymes it has been proposed that hyperthermophilic archaea utilize the non-phosphorylative ED pathway for glucose degradation[2,5-10]. However, using a comparative genomics approach we detected a conserved ED cluster in the genomes of *T. tenax*[2], *S. solfataricus*, and *S. tokodaii* that resembles the cluster present in *Halobacterium* sp. NRC1 (Figure 8.2).

The ED gene cluster

The ED gene clusters of the hyperthermophilic crenarchaea *T. tenax*[2] and *S. solfataricus* comprise genes encoding KD(P)G aldolase (COG0329), a putative KDG kinase (COG0524) and a putative gluconate dehydratase (GAD, *gad* gene, COG4948). In addition, a gene encoding a glucan-1,4-α-glucosidase (GAA, *gaa* gene, COG3387) was identified in the *T. tenax* ED cluster, and a gene coding for a non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN, *gap*N gene, COG1012) was found in the *S. solfataricus* ED gene cluster (Figure 8.2). The *S. tokodaii* gene cluster resembles that of *S. solfataricus*, however in this organism, the *gad* gene is not part of the gene cluster. Homologs of the *gad* and *kdg*A gene (both share 29 % identity to the *T. tenax* enzymes) are also present in the ED cluster of *Halobacterium* sp. NRC1 (*gcd*, *orf-kdg*A-*gad*), which additionally comprises genes that encode glucose dehydrogenase (*gcd*, COG1063) and a hypothetical protein (Figure 8.2).

Whereas the *S. solfataricus* KDG aldolase was cloned and characterized previously[21,22,25], no functional information was available on the other genes present in the gene cluster when this study was initiated[2-4]. The genes encoding the putative dehydratases (COG4948) that reside in the *T. tenax* and *S. solfataricus* ED clusters, exhibit high similarity to members of the diverse mandelate racemase/muconate lactonizing enzyme (MR-MLE) subgroup of the enolase superfamily, which includes some sugar dehydratases[26]. The conserved clustering of these genes within the ED clusters of *T. tenax* and *S. solfataricus*, suggests that their gene products are likely candidates for the missing gluconate dehydratase (GAD, *gad* gene) activity (Figure 8.2). Despite the fact that GAD activity has been measured in different bacterial, archaeal and eukaryal (fungal) sources[27-30], the gene has not been identified for a long time. As such, it represents a missing link in central carbohydrate metabolism, and only recently GAD activity was reported for the recombinant *T. tenax*[2] and the purified *S. solfataricus* enzymes[31,32]. The enzyme shows no homology to the classical phosphogluconate dehydratase (ED dehydratase, *edd*; EC 4.2.1.12, COG0129).

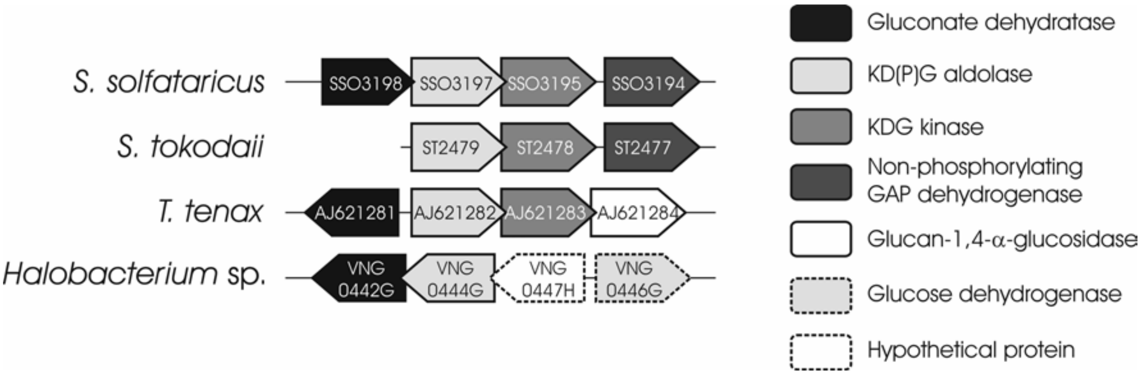


Figure 8.2 Schematic representation of the conserved gene clusters in archaeal genomes, comprising key genes of the semi-phosphorylative ED pathway. The genes are indicated by their systematic gene name; except for *T. tenax*, the accession numbers are displayed and orthologous genes are shaded in the same grey scale. Genes are not drawn to scale.

The conserved clustering of genes encoding putative KDG kinases in addition to *gap*N (in *S. solfataricus* and *S. tokodaii*) was rather surprising, since the modified ED pathway has been reported to proceed via non-phophorylated intermediates at the C-6 position. This conserved gene clustering did suggest the operation of the semi-phosphorylative pathway in these organisms. Moreover, the high similarity of the thermophilic proteins to the haloarchaeal KDG kinase and KDPG aldolase suggests similar substrate specificity, again indicating the presence of the semi-phosphorylative ED pathway in *T. tenax* and *S. solfataricus*.

Enzyme characterization

The *gdh*, *gad*, *kdg*A, and *kdg*K gene of *T. tenax* and the *gad*, *kdg*A, *kdg*K, and *gap*N genes of *S. solfataricus* were cloned and expressed in *E. coli* using the pET expression system. The recombinant enzymes were enriched from crude extracts by heat precipitation (GDH of *T. tenax*; GAD, KDG kinase and GAPN of *S. solfataricus*) and key enzymes were further purified by gel filtration (GAD, KD(P)G aldolase, KDG kinase of *T. tenax*; KD(P)G aldolase of

S. solfataricus. Due to the heat precipitation at 65-90°C (30 min) and the high assay temperature at 70°C the activity of residual, contaminant *E. coli* proteins is very unlikely and was further diminished by analysis of a heat precipitated extract of the expression host with plasmid without insert.

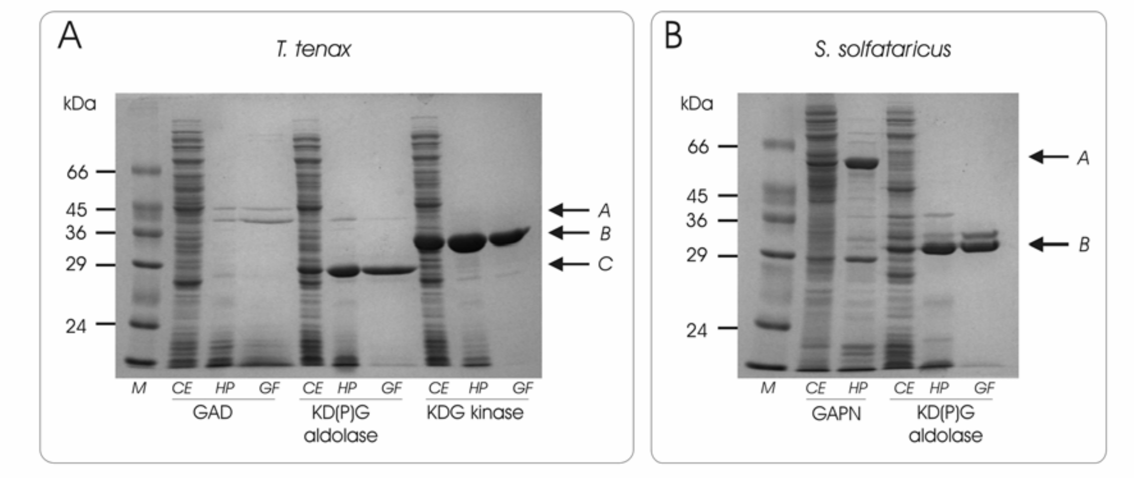


Figure 8.3 (a) SDS/PAGE of recombinant expression and purification of *T. tenax* GAD, KD(P)G aldolase and KDG kinase. Arrows indicate the purified recombinant GAD (*A*), KD(P)G aldolase (*C*) and KDG kinase (*B*). **(b)** SDS/PAGE of recombinant expression and purification of *S. solfataricus* KD(P)G aldolase and GAPN. Arrows indicate the purified recombinant GAPN (*A*) and KD(P)G aldolase (*B*). Lanes containing crude cell extracts (CE), soluble fractions after heat precipitation (HP) and gel filtration (GF) were loaded with 20, 10 and 5 µg of protein respectively. Lane M corresponds to the protein marker, Dalton Mark VII-L (Sigma).

For all enzymes, with the exception of the *S. solfataricus* GAD and KDG kinase, a good expression and sufficient enrichment was observed (Figure 8.3). For the *T. tenax* GAD, two bands were enriched after gel filtration (Figure 8.3A); however, both proteins exhibited different elution profiles after gel filtration without salt and the upper band was clearly associated with catalytic GAD activity. Whereas, as observed from SDS/PAGE, the molecular mass approximately corresponds to the calculated mass for *T. tenax* GAD (43/44.033 kDa) and *T. tenax* KDG kinase (32/ 33.308 kDa), some deviation is observed for both KD(P)G aldolases [*T. tenax* (26/30.982 kDa), *S. solfataricus* (30/33.108)] and *S. solfataricus* GAPN (54/56.927) (apparent/calculated molecular masses are given in parentheses). However, these differences of 3–5 kDa are in agreement with generally observed minor deviations.

Expression of *S. solfataricus* GAD and KDG kinase was rather poor, as only little recombinant protein was observed in the soluble fraction. In addition, unlike the native enzyme[31,32], the recombinant *S. solfataricus* GAD appeared to be relatively instable, not allowing the heat precipitation step to be performed above 65 °C. Attempts to improve the poor recombinant expression by the use of different expression hosts [BL21 (DE3), BL21 (DE3) CodonPlus, JM109 (DE3) and ROSETTA] and different suspension buffers were not successful.

GDH

The *T. tenax* glucose dehydrogenase (GDH) catalyzes the oxidation of glucose yielding gluconate. The gene was unequivocally identified by the previously determined N-terminal sequence of the GDH isolated and characterized from *T. tenax* cells and confirmed by the activity of the recombinant protein (data not shown)[10]. The enzyme was used for the *in vitro* reconstruction of the pathway.

GAD

Gluconate dehydratase (GAD) catalyzes the dehydration of gluconate yielding KDG. The *T. tenax* GAD was assayed by monitoring KDG formation using the discontinuous TBA assay (70°C). The enriched *T. tenax* enzyme after heat precipitation (0.272 ±0.007 U/mg protein) exhibits high sensitivity towards salts (KCl). After ion chromatography (Q-sepharose, Amersham Biosciences) all activity was lost and after gelfiltration in the presence of 200 mM KCl the activity is significantly reduced. As shown in Figure 8.4 the time-dependent formation of KDG from gluconate was only observed in the presence of enzyme (fraction after gelfiltration) and substrate (*T. tenax*: 0.065 ±0.006 U/mg protein at 10 mM gluconate), thus confirming the proposed GAD activity.

The specific activity is proportional to the amount of recombinant protein (0.0699 ±0.003 U/mg (30 μg); 0.061 ±0.003 U/mg (60 μg). No activity could be detected with negative controls without protein (gluconate, galactonate), without substrate (data not shown), as well as with a heat-precipitated cell-free extract of BL21 (DE3) CodonPlus with plasmid pET-15b without insert. Considering the reports of pathway promiscuity in *S. solfataricus*[22] the activity of the *T. tenax* GAD with galactonate as substrate was analyzed, but no activity was observed in the protein fraction after heat precipitation or gel filtration. In agreement with this finding GAD activity on gluconate and not galactonate was observed in crude extracts of *T. tenax*, whereas activity with both substrates was observed in *S. solfataricus* crude extracts (Table II).

The *S. solfataricus* GAD exhibits some activity at 65°C, but its high thermal instability did not allow further characterization. While this manuscript was submitted for publication, the purification and characterization of the GAD from *S. solfataricus* was reported by two independent groups[31,32]. Surprisingly, these studies report some contradicting results about molecular size and catalytic activity with galactonate; our analysis supports the proposed promiscuity in *S. solfataricus*[32].

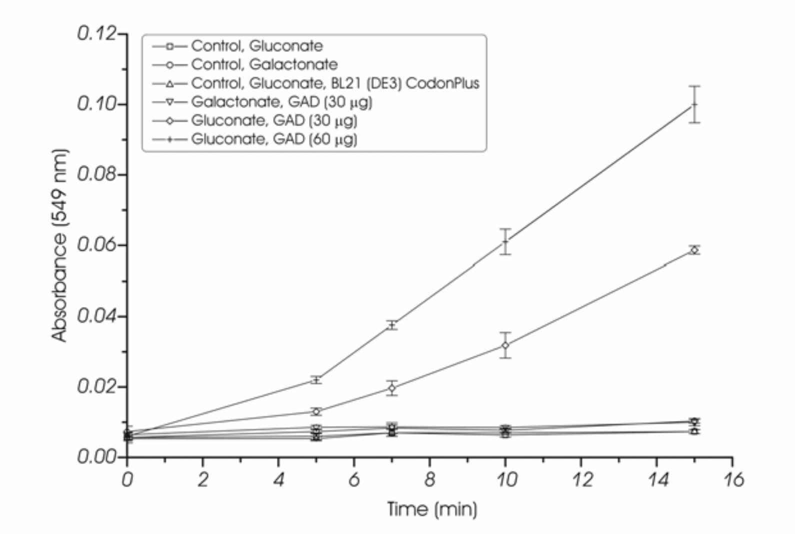


Figure 8.4 *T. tenax* GAD activity (protein fraction after gel filtration) was monitored at 70°C using the discontinuous TBA assay. Activity on gluconate (10 mM; 30 and 60 µg of protein) as well as galactonate (10 mM; 30 µg of protein) and controls without enzyme (10 mM gluconate and galactonate) or substrate (results not shown) and heat-precipitated extract of BL21 (DE3) CodonPlus with pET-15b without insert are shown. All experiments were performed in triplicate and the S.D. is given. GAD activity was only observed in the presence of gluconate and the observed activity is proportional to the amount of enzyme.

The GADs of *T. tenax* and *S. solfataricus* are members of the mandelate racemase (MR) subgroup of the enolase superfamily, and thus represents the first reported GAD in this superfamily[26]. The enolase superfamily harbors a diverse set of enzymes, which at first sight catalyze markedly different overall reactions (e.g. enolase, mandelate racemase, galactonate dehydratase, muconate-lactonizing enzyme I, β-methylaspartate ammonia lyase, o-succinylbenzoate synthase). A common feature of all family members concerns the first step of their catalytic action, i.e. the abstraction of the α-proton of a carboxylic acid to form an enolic intermediate. The enolase superfamily is divided in the MR, muconate-lactonizing enzyme I (MLE I) and enolase subgroup. In the MR subgroup so far only glucarate dehydratase from *Pseudomonas putida*, *Bacillus subtilis* and *E. coli* as well as galactonate dehydratase from *E. coli* have been biochemically characterized[26,33]. Homologs of the *S. solfataricus* and *T. tenax* GAD have been identified in many archaeal, bacterial and eukaryal species.

However, functional annotation of these proteins remains difficult due to the broad substrate specificity within the enolase family.

KDG aldolase

KDG aldolase catalyzes the reversible cleavage of KDG yielding pyruvate and glyceraldehyde (GA). The *T. tenax* and *S. solfataricus* KDG aldolase activity was assayed in the anabolic direction of KDG formation from C-3 substrates (condensation reaction) using the discontinuous TBA assay. In contrast to previous reports on the *S. solfataricus* KDG aldolase[21] activity was observed not only with GA but also with the phosphorylated substrate GAP. The time-dependent formation of KDG and KDPG was monitored in the presence of the *T. tenax* or *S. solfataricus* enzyme and GA and GAP, respectively, as substrate (Figure 8.5). The observed activity is proportional to the amount of recombinant protein, as shown for GA (*T. tenax*: 2.091 ±0.086 U/mg protein (6 μg), 2.051 ±0.107 U/mg protein (12 μg); *S. solfataricus*: 2.186 ±0.255 U/mg protein (6 μg), 1.945 ±0.129 U/mg protein (12 μg)) and GAP (*T. tenax*: 11.158 ±0.442 U/mg protein (6 μg), 11.774 ±0.532 U/mg protein (12 μg); *S. solfataricus*: 5.296 ±0.855 U/mg protein (6 μg), 5.558 ±0.356 U/mg protein (12 μg)). The negative controls without protein, only one substrate, and cell-free extract of expression host with empty vector (results not shown) revealed no activity.

In order to confirm these results both enzymes were assayed in the presence of ^{14}C-labelled pyruvate, the products were separated by TLC and afterwards monitored by autoradiography (Figure 8.6). In agreement with the aforementioned enzyme assays, the formation of both labelled products was observed: KDG from GA and pyruvate, or KDPG from GAP and pyruvate. No product formation was observed in the controls without protein, only one substrate (GA, GAP or pyruvate, respectively) and with cell-free extract of the host BL21 (DE3) with plasmid pET-15b without insert after heat precipitation. As an additional control the characterized KDPG aldolase (EDA) of the anaerobic, hyperthermophilic bacterium *T. maritima*, which was reported for activity on phosphorylated and non-phosphorylated substrates[34] was used. In accordance to the *T. tenax* and *S. solfataricus* KD(P)G aldolase KDG and KDPG formation was observed. (Figure 8.6) The different spots of pyruvate on TLC represent the dissociated and un-dissociated form. The observed change in the pattern of the labelled pyruvate (second spot) is obviously due to the presence of GAP alone, as shown by the controls without KD(P)G aldolase and BL21 (DE3) extract.

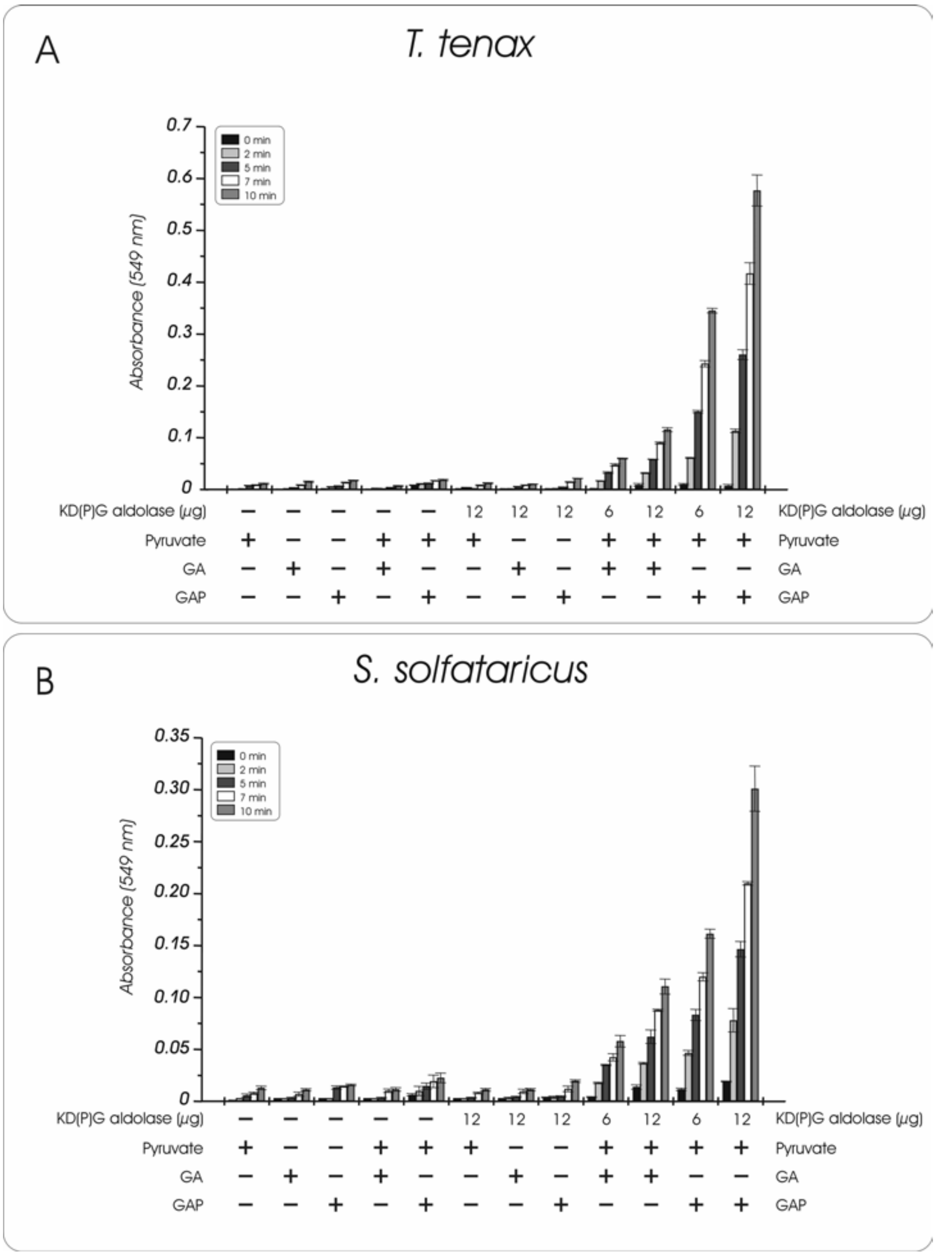


Figure 8.5 KD(P)G aldolase activity of *T. tenax* **(a)** and *S. solfataricus* **(b)**. The formation of KDG and KDPG from pyruvate (5 mM) and GA or GAP (2 mM) respectively was monitored at 70°C using the discontinuous TBA assay. The dependence on the amount of protein (6 and 12 µg of protein, fraction after gelfiltration) and controls with one (GA, GAP or pyruvate) or both (pyruvate and GA or GAP respectively) substrates without enzyme and with one substrate (GA, GAP or pyruvate respectively) in the presence of enzyme are shown. For each probe, three independent measurements were performed and the experimental error is given. In the presence of KD(P)G aldolase, the formation of KDG from GA and pyruvate as well as KDPG from GAP and pyruvate was observed. The activity with non-phosphorylated and phosphorylated substrates is proportional to the amount of enzyme used in the assay.

Thus, in contrast to former reports, both the *T. tenax* and *S. solfataricus* enzyme are 2-keto-3-deoxy-(6-phospho)-gluconate (KD(P)G) aldolases of low substrate specificity that are active on phosphorylated (GAP, KDPG) as well as non-phosphorylated (GA, KDG) substrates. The recombinant *S. solfataricus* aldolase was re-investigated in the group of Michael Danson (University of Bath, Bath, UK) and the activity on phosphorylated substrates was indeed confirmed (M. Danson personal communication). This bi-functional KD(P)G aldolase appears to be a key enzyme in both the non- and the semi-phosphorylative ED pathway, which would be in line with the presence of the *kdg*A gene homolog in the haloarchaeal ED cluster.

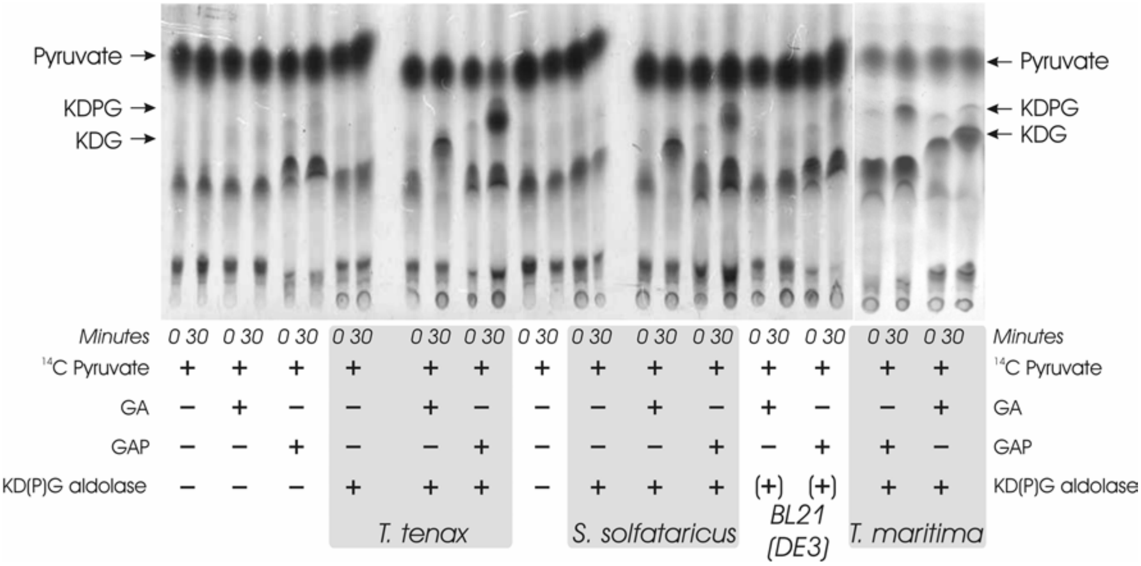

Pyruvate
KDPG
KDG
Pyruvate
KDPG
KDG
Minutes 0 30 0 30 0 30 0 30 0 30 0 30 0 30 0 30 0 30 0 30 0 30 0 30 0 30 0 30 Minutes
¹⁴C Pyruvate + + + + + + + + + + + + + + ¹⁴C Pyruvate
GA − + − − + − − − + − + − − + GA
GAP − − + − − + − − − + − + + − GAP
KD(P)G aldolase − − − + + + − + + + (+) (+) + + KD(P)G aldolase
T. tenax
S. solfataricus
BL21 (DE3)
T. maritima

Figure 8.6 Detection of ^{14}C-labelled KDG and KDPG through TLC and autoradiography. The KD(P)G aldolases of *T. tenax* and *S. solfataricus* (fraction after heat precipitation) were incubated at 70°C in the presence of labelled pyruvate ([2-^{14}C]pyruvate) and either GA or GAP. Control samples containing different combinations of KD(P)G aldolase and substrates are shown as indicated. In addition, control samples of the expression host BL21 (DE3) CodonPlus with pET-15b without insert after heat precipitation, indicated by '+', and formation of ^{14}C-labelled KDG and KDPG by the KDPG aldolase (EDA) of *T. maritima* (fraction after heat precipitation) [34] are shown. The formation of ^{14}C-labelled KDG and KDPG was monitored via TLC and visualized using autoradiography. As for the KDPG aldolase of *T. maritima* , the formation of both KDG (from GA and pyruvate) and KDPG (from GAP and pyruvate) is observed in the presence of the KD(P)G aldolase of *T. tenax* or *S. solfataricus*.

Recent studies of the *S. solfataricus* KD(P)G aldolase revealed a lack of facial selectivity in the aldolase reaction, catalyzing the formation as well as the cleavage of both KDG and KDgalactonate, with glyceraldehyde and pyruvate as substrate and product, respectively[22]. At present no information is available about facial selectivity for the *T. tenax* enzyme.

Homologs of the *T. tenax* and *S. solfataricus* KDG aldolases were identified in many bacterial, archaeal and eukaryal species. However, they share no similarity to the classical ED aldolase (EDA)[21], but are members of the N-Acetylneuraminate lyase (NAL) superfamily[35]. Although members of the NAL superfamily catalyze substantially different overall reactions (e.g. dihydrodipicolinate synthases, N-acetylneuraminate lyase, trans-o-hydroxybenzylidene-pyruvate aldolase/dehydratase), their catalysis generally proceeds via a Schiff base mechanism. Each of the enzymes of the NAL superfamily harbors a conserved lysine (Lys-165 in NAL) located in the sixth strand of β-sheet of the single β/α (TIM) barrel domain[35]. The corresponding active site lysine residue has indeed been identified in the crystal structure of the KD(P)G aldolase of *S. solfataricus* (Lys-155)[36]. In *E. coli* K12 two KD(P)G aldolase homologs in addition to the classical EDA were identified (*yjh*H, *yag*E). Both are organized in gene clusters encoding ED dehydratase orthologs (*yjh*G, *yag*F[37]), permeases (*yjh*F, *yag*G), regulators (*yjh*I, *yag*I) and a hypothetical protein (*yjh*U) or a putative β-xylosidase (*yag*H), respectively (not shown). This functional organization indicates that also in *E. coli* related, so far unknown ED modifications may exist.

KDG kinase

KDG kinase catalyzes the phosphorylation of KDG yielding KDPG. The activity of the KDG kinases of *T. tenax* and *S. solfataricus* was only observed in the presence of ATP and Mg^{2+}. The KDG kinase activity of *T. tenax* (fraction after gel filtration) was followed in response to different substrate concentrations, and as shown in Figure 8.7, the enzyme follows Michaelis-Menten kinetics for KDG (K_m of 0.178 ±0.0114 mM, V^{max} of 43.3 ±0.007 U/mg protein). The measured enzyme activity of *T. tenax* was directly proportional to the amount of enzyme added to the assay ((41.88 ±0.23 U/mg (1.5 μg), 41.32 ±0.13 U/mg (3 μg)), at 5 mM KDG). Since the expression of the *S. solfataricus* KDG kinase was rather poor the activity was determined directly after heat precipitation in the presence of 3 mM KDG and was shown to be directly proportional to the amount of enzyme added to the assay (0.103 ±0.004 U/mg (40 μg) and 0.104 ±0.014 U/mg (80 μg)) (Table I).

The coupled KDG kinase assay is not optimal, since KD(P)G aldolase is also active on the substrate KDG, resulting in an unknown effective KDG concentration in the assay. Furthermore, the *S. solfataricus* KD(P)G aldolase was reported to form the diastereomeric products KDG and KDgalactose by condensation of glyceraldehyde and pyruvate. Therefore, the same assay was tested by generating KDG from gluconate by the *T. tenax* GAD during the assay, which forms specifically KDG, but again allows no determination of the effective KDG concentration. The measured enzyme activity was directly dependent on the gluconate concentration (6.57 U/mg ±0.46, 12.18 ±0.60, 14.81 ±0.30 and 22.47 ±0.93 U/mg protein with 1, 1.5, 5 and 10mM gluconate, respectively).

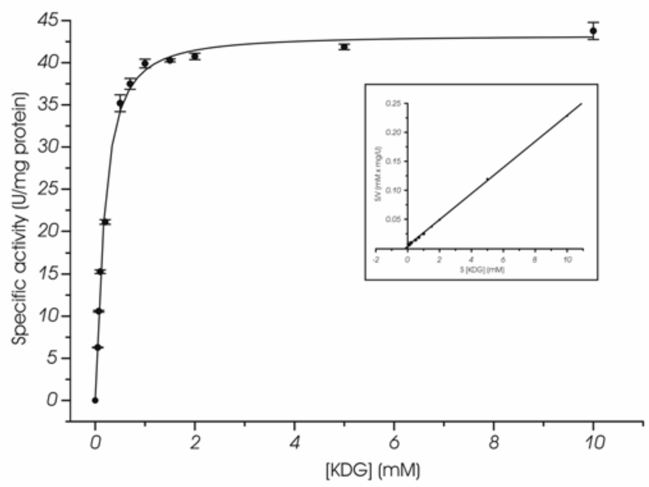


Figure 8.7 KDG kinase activity of *T. tenax*. The KDG kinase activity was determined in a continuous assay at 70°C by monitoring the formation of GAP after KDPG cleavage via KD(P)G aldolase and GAPN of *T. tenax*. The rate dependence on the KDG concentration, determined by the TBA assay, is shown. The enzyme follows Michaelis–Menten kinetics for KDG. The inset shows the linear transformation according to Hanes. For the assay, it was ensured that the amount of auxiliary enzymes is not rate-limiting and the measured enzyme activity was directly proportional to the amount of enzyme added to the assay (results not shown). Three independent assays were performed for each substrate concentration and the S.D. is given. The rate-dependent formation of GAP was only monitored in the presence of KDG, ATP, Mg^{2+}, auxiliary enzymes and the KDG kinase of *T. tenax*.

In addition, the KDG kinase activity of both enzymes was also analyzed using a discontinuous assay by monitoring the formation of ADP from the ATP-dependent phosphorylation of KDG (generated by GAD) via pyruvate kinase and lactate dehydrogenase. The time-dependent formation of ADP was only observed in the presence of gluconate, GAD and the KDG kinase of *T. tenax* or *S. solfataricus*, whereas no ADP formation was observed with the negative controls (no protein, GAD alone, KDG kinase without GAD, cell free extract with empty vector) (data not shown).

Obviously, the KDG kinase is the key enzyme in the semi-phosphorylative ED pathway. The enzyme is a member of the ribokinase (PfkB) enzyme family, which is composed of prokaryotic sequences related to ribokinase, including enzymes such as fructokinases, the minor 6-phosphofructokinase of *E. coli*, 1-phosphofructokinase and archaeal ADP-dependent glucokinases and phosphofructokinases[38,39]. So far the KDG kinase purified from *E. coli* was characterized[40] and activity was demonstrated for the gene product of *Erwinia chrysanthemi*[41] (accession number X75047, 25% identity to the *T. tenax* enzyme). The latter enzyme is involved in pectin and hexuronate (glucuronate and galacturonate) catabolism, routes that converge through the common intermediate KDG. KDG kinase activity has also been proposed for the *kdg*K gene of *Bacillus stearothermophilus* T6[42] due to its high similarity to the *Erwinia* enzyme and the organization in the xylan and glucuronic acid utilization gene cluster.

Homologs of the *T. tenax* and *S. solfataricus* KDG kinase were identified in many archaeal genomes (e.g. A. pernix (28% identity), P. furiosus (30% identity), *P. aerophilum* (31% identity)) and bacterial genomes (*Pseudomonas putida* (34% identity), *Bacillus halodurans* (30% identity), *Streptomyces coelicolor* (33% identity)). Interestingly, KDG kinase orthologs were not detected in the genomes of *T. acidophilum*, *T. volcanium* and *Picrophilus torridus*. This suggests either that a strictly non-phosphorylative ED pathway is operative in these thermophilic archaea, or that the *kdg*K gene has been substituted via a non-orthologous gene displacement. In addition, we failed to identify KDG kinase orthologs in the Eukarya and thus, the KDG kinase seems to be a key player in glucose catabolism (semi-phosphorylative ED pathway) and sugar acid (extracellular polymer) degradation in prokaryotes.

GAPN

GAPN catalyzes the irreversible, non-phosphorylating oxidation of GAP to 3-phophoglycerate (Figure 8.1). The *T. tenax* GAPN has been studied in detail previously[43-45]. The GAPN of *S. solfataricus* shows high similarity (56% identity) to the enzyme of *T. tenax*. The clustering of *gap*N with ED genes in *S. solfataricus* and *S. tokodaii* underlines the role of GAPN in the semi-phosphorylative ED pathway and more general in the common shunt of the EMP pathway.

The *S. solfataricus* GAPN activity was determined in a continuous assay at 70°C monitoring the formation of NADPH or NADH. The enzyme follows Michaelis Menten kinetics for $NADP^+$ and GAP. For $NADP^+$ a K_m of

0.099 ±0.009 mM and V^{max} of 3.28 ±0.102 U/mg protein and for GAP a K_m of 0.453 ±0.052 mM and a V^{max} of 3.36 ±0.146 U/mg protein was determined (Table I). GAP concentrations above 4 mM showed an inhibitory effect on GAPN activity. The enzyme shows only remote activity with NAD^+ and no saturation of the enzyme was observed up to 10 mM NAD^+ (results not shown). A detailed characterization of the *S. solfataricus* enzyme is currently underway (T.J.G. Ettema, H. Ahmed, B. Siebers, J. van der Oost, unpublished work).

In summary, the proposed activities of the ED genes could be confirmed by analysis of the recombinant gene products: GDH, GAD, KD(P)GA, KDGK and GAPN (Table I). In addition the dual activity of the KD(P)G aldolase, which is active on non-phosphorylated as well as phosphorylated substrates, suggests that in contrast to previous assumptions, the semi-phosphorylative ED pathway is operative in these hyperthermophiles.

Table I. Specific activities of the recombinant ED enzymes of *T. tenax* and *S. solfataricus*

Enzyme activity	Substrate concentration	Specific activity (U/mg)
T. tenax		
GAD	10 mM Gluconate	0.065±0.006
GAD	10 mM Galactonate	n.d.
KD(P)G aldolase	2 mM Glyceraldehyde 5 mM Pyruvate	2.1±0.1
KD(P)G aldolase	2 mM Glyceraldehyde 3-P 5 mM Pyruvate	11.5±0.6
KDG kinase	5 mM Gluconate 2 mM ATP	41.6±0.3
S. solfataricus		
KD(P)G aldolase	2 mM Glyceraldehyde 5 mM Pyruvate	2.2±0.2
KD(P)G aldolase	2 mM Glyceraldehyde 3-P 5 mM Pyruvate	5.8±0.7
KDG kinase[a]	5 mM Gluconate[b] 2 mM ATP	0.104±0.001
GAPN[a]	3 mM Glyceraldehyde 3-P 2 mM $NADP^+$	3.36±0.08

[a] Protein fraction after heat precipitation.

[b] KDG was generated by KD(P)G aldolase and purified by anion-exchange chromatography.

In vitro reconstruction of the ED pathways

In order to analyse the activities of the different ED enzymes and to confirm their function in the ED pathway U-^{14}C glucose was incubated in presence of different combinations of ED enzymes (glucose dehydrogenase

(*T. tenax* GDH), gluconate dehydratase (*T. tenax* GAD), KD(P)G aldolase (*T. tenax* KD(P)GA), KDG kinase (*T. tenax* KDGK)) and co-substrates ($NADP^+$, ATP, Mg^{2+}). Subsequently, the labelled intermediates that were formed during the incubations, were separated by TLC and afterwards detected by autoradiography (Figure 8.8).

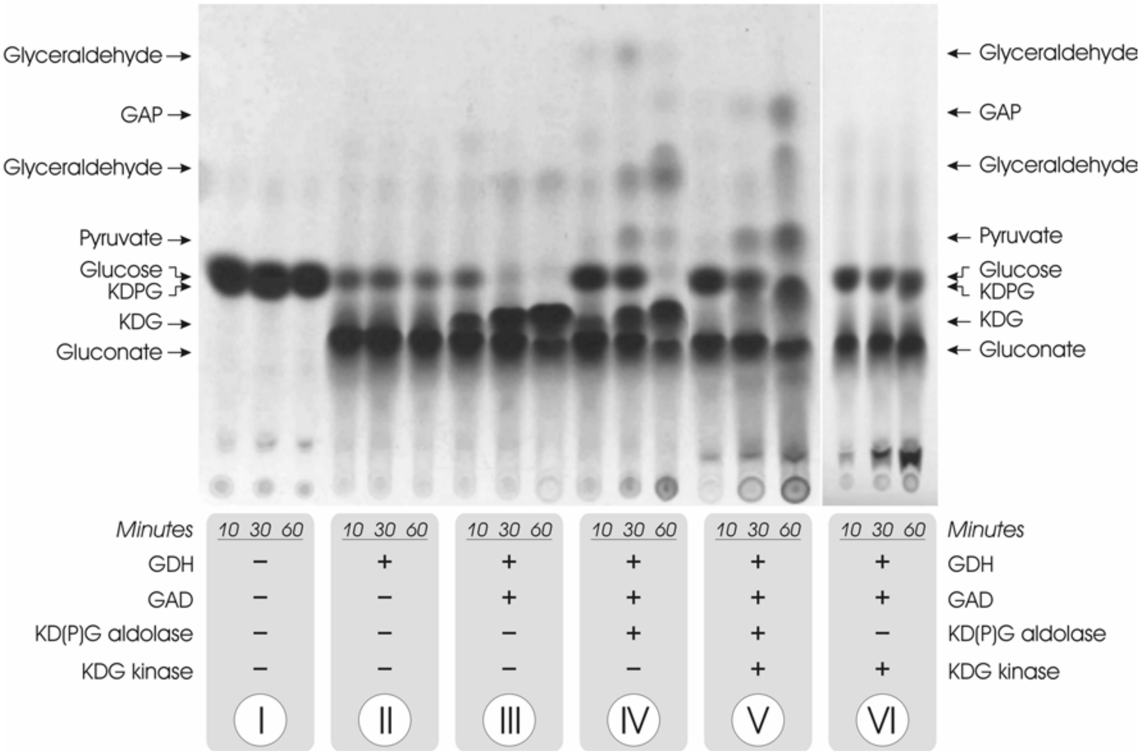


Figure 8.8 Reconstruction of the ED pathway *in vitro*. [U-^{14}C]glucose was incubated in the presence of different combinations of ED enzymes from *T. tenax* [GDH, GAD, KD(P)G aldolase and KDG kinase; protein fractions after heat precipitation] as indicated (10, 30 and 60 min at 70°C respectively), and the labelled intermediates were separated by TLC and visualized using autoradiography. The labelling pattern in the presence of KDG kinase and KD(P)G aldolase (V) indicates the co-existence of both the semi-phosphorylative and non-phosphorylative ED modifications in *T. tenax*.

The step-wise addition of GDH, GAD and KD(P)G aldolase to labelled glucose (I-IV, Figure 8.8) reveals the characteristic intermediates of the non-phosphorylative ED pathway: gluconate, KDG, pyruvate and glyceraldehyde. However, after the addition of KDG kinase and co-substrates (ATP, Mg^{2+}) KDPG formation is observed in the presence or the absence of KD(P)G aldolase (V, VI; Figure 8.8) while KDG has disappeared. In addition, in the presence of KD(P)G aldolase formation of GAP is observed, as characteristic intermediate of the semi-phosphorylative ED pathway, in addition to formation of gluconate, pyruvate and some glyceraldehyde. The identification of GA and GAP in this sample indicates that, at least *in vitro*, both the non-phosphorylative and the semi-phosphorylative versions of the ED pathway are active in parallel. Identical labelling patterns were observed when using the KD(P)G aldolase and KDG kinase of *S. solfataricus* instead of the two *T. tenax* enzymes (data not shown).

A possible role of the ED pathway in gluconeogenesis was analyzed using a similar approach with ^{14}C-labelled pyruvate. However, in the presence of KD(P)G aldolase, GAD and GDH and the respective substrates and co-substrate (GA, pyruvate, NADPH) only the formation of KDG was observed (data not shown), indicating that the pathway is at least partly irreversible, or it is catalyzed by a distinct set of enzymes.

In addition, the *in vitro* reconstruction experiments demonstrate that gluconolactonase activity (EC 3.1.1.17) is not needed for the functional reconstruction of the pathway. Interestingly, whereas a potential gluconolactonase encoding gene seems to be absent in the *T. tenax* genome, potential candidates (SSO3041 and ST2555) were identified adjacent to the glucose dehydrogenase encoding genes (SSO3042 and ST2556) in *S. solfataricus* and *S. tokodaii*. The clustering of these genes in these organisms indicates a functional link between their gene products. Possibly, the presence of a gluconolactonase encoding gene allows an accelerated glucose turnover in *S. solfataricus* and *S. tokodaii*, since the ED pathway seems to be the only route for glucose catabolism in these organisms.

In vivo operation of a branched ED Pathway in *T. tenax* and *S. solfataricus*

The aforementioned *in vitro* results strongly suggest functional non- and semi-phosphorylative versions of the ED pathway in *T. tenax* and *S. solfataricus* (Figure 8.1). This would be the first demonstration of the presence of the semi-phosphorylative ED pathway in *T. tenax* and *S. solfataricus* in particular, and in hyperthermophilic archaea in general, and the first indication of the presence of both ED modifications in one organism (pathway dualism). The conserved functional organization of ED genes encoding enzymes of the semi-phosphorylative modification (KDG kinase, GAPN) and for the common, ”core” modified ED shunt (GAD; KD(P)G aldolase) raises questions about their regulation as well as the utilization of both modified ED versions *in vivo*.

In order to prove that the two ED modifications are operative *in vivo*, enzyme assays were performed on crude extracts of *T. tenax* and *S. solfataricus* cells grown on glucose. As shown in Table II, the enzyme activities of GAD, KD(P)G aldolase, KDG kinase and GAPN could be measured in these crude extracts, indicating that the operation of a semi-phosphorylative ED pathway is possible in these organisms. As a peculiarity GAD activity on galactonate was only observed in *S. solfataricus* crude extracts supporting our studies of the recombinant *T. tenax* enzyme and reports about pathway promiscuity in *S. solfataricus*. In addition, it is demonstrated that aldolase activity could be demonstrated with pyruvate and either GA or GAP as substrates. This observation is in good agreement with the results of the characterization of the recombinant KD(P)G aldolase. However, our results do not rule out that a non-phosphorylative ED pathway is operating in parallel with the semi-phosphorylative version.

At present no information is available for the specific enzymes of the non-phosphorylative ED pathway. Several dehydrogenases and aldehyde oxidoreductases as well as a homolog for glycerate kinase (COG2379, *T. tenax* AJ621345, SSO0666) are encoded by the *T. tenax* and *S. solfataricus*[2-4] genomes, allowing the operation of the non-phosphorylative ED modification in both organisms. The putative glycerate kinase of *T. tenax* shows high similarity to the orthologs identified in *S. solfataricus* (26 % identity), T. acidophilum (23 % identity) and Thermotaga maritima (33 % identity), however so far no functional information is available for one of the gene products. Since glycerate kinase represents the key activity of the non-phosphorylative ED pathway, the expression of the *T. tenax* enzyme is on the way in order to elucidate its function *in vivo*.

Table II. Specific activities ED enzymes in crude extracts of *T. tenax* and *S. solfataricus*

Enzyme activity	**Substrate concentration**	**Specific activity (mU/mg)**	
		T. tenax	***S. solfataricus***
GAD	10 mM Gluconate	10.9±0.7	15.2±0.9
	10 mM Galactonate	n.d.	5.7±0.6
KD(P)G aldolase	2 mM GA, 5 mM Pyruvate	1.4±0.1	1.5±0.4
	2 mM GAP, 5 mM Pyruvate	1.7±0.3	13.4±0.8
KDG kinase	10 mM KDG, 2 mM ATP	9.1±0.5	8.7±0.3
GAPN[a]	3 mM GAP, 2 mM $NADP^+$	1.13±0.04[a]	32.3±0.5
	3 mM GAP, 1 mM NAD^+	3.5±0.1	n.d.[a]

[a] Assay was performed in the presence of 20 mM $NADP^+/NAD^+$.

Physiological implications

The presence of various pathways for carbon metabolism in one organism raises questions about their physiological significance. With the anaerobe *T. tenax* and the aerobe *S. solfataricus* two hyperthermophiles are studied, which exhibit significant differences in the central carbon metabolism and thus might allow gaining new insights into the flexibility of carbon metabolism.

Genome data and biochemical studies indicate that the anaerobe *T. tenax* uses at least two different pathways for glucose metabolism, a modification of the reversible EMP pathway and one or two modifications of the ED pathway, a semi-phosphorylative and a non-phosphorylative version. The clustering of the *gaa* gene, encoding a glucan-1,4-α-glucosidase, with the ED genes indicates a central role of the ED modifications in the hydrolytic degradation of polysaccharides (e.g. glycogen). In contrast, the modified EMP pathway seems to have a central function in the phosphorolytic glycogen degradation by glycogen phosphorylase, which was characterized recently[2,46]. The selection of the different pathways *in vivo* seems to be strongly influenced by the energy demand of the cell. Whereas no ATP is generated by the ED modifications, one (glucose degradation) or two ATP (phosphorolytic glycogen degradation) are generated using the EMP variant, taking into account that: (i) PP_i, the phosphoryl donor of phosphofructokinase, is a waste product of the cell[47], and (ii) GAPN is used for glucose catabolism, which omits the formation of 1,3-diphosphoglycerate and as such does not couple the oxidation of GAP to the generation of ATP.

In the aerobe *S. solfataricus* the modifications of the ED pathway seem to represent the only pathway for glucose and galactose degradation[22]. Analyses of genome data indicate an incomplete EMP pathway, and it is suggested that the enzymes that are present may be involved in fructose degradation or in the anabolic gluconeogenetic direction for glycogen synthesis[3,4,48].

The identification and characterization of the enzymes that constitute the modified ED pathway sheds new light in the functional role of this glycolytic pathway in hyperthermophilic archaea and suggests a much broader distribution of ED-like pathways in other Archaea, Bacteria and Eukarya than was previously assumed. This finding supports the important role of the ED pathway and its variants in glucose degradation and as a funnel for sugar acid (polymer) degradation and again underlines the variability and flexibility of central carbon metabolizing pathways.

EXPERIMENTAL PROCEDURES

Strains and growth conditions

Cultures of *T. tenax* (DSM 2078[23]) and *S. solfataricus* P2 (DSM 1617[49]) were grown as reported previously[50,51]. For *S. solfataricus* carbon sources were added to a final concentration of 0.2% (w/v). *E. coli* strains DH5α (Life Technologies), BL21 (DE3), BL21 (DE3) CodonPlus (Novagen), and JM109 (DE3) (Promega) for cloning and expression studies were grown under standard conditions[52] following the instructions of the manufacturer.

Table III. **Used primer sets, plasmids and hosts**

Genes	S/A[a]	RE[a]	Sequence (5'→3')[b]	Plasmid	Host
T. tenax					
gdh	S	*Bsp*HI	TAGAGGCT**C**ATGAGGGCTG	pET-15b	BL21(DE3)CP
	A	*Bam*HI	ACTACCGTG**GAT**CCACAAC		
gad	S	*Nco*I	TTTGGCCAGCGC**C**ATGGCCTCATCG	pET-15b	BL21(DE3)CP
	A	*Xho*I	AAATGCCGGC**C**TCGA**G**GGAATGGGA		
*kdg*A	S	*Nco*I	AGGGCGCCCCGAGTACTATC**C**ATGGAGA	pET-15b	BL21(DE3)CP
	A	*Xho*I	GGGGCTCCCCT**C**G**A**GCTACCAGGC		
*kdg*K	S	*Nde*I	GAGCCAGCTGAG**CA**TATGATAAGCCTGG	pET-15b	BL21(DE3)CP
	A	*Eco*RI	TTGCCCAGAAT**TC**CGCTCCTC		
gaa	S	*Nde*I	TATTGAAGGC**CA**TATGAGGAG	pET-15b	BL21(DE3)CP
	A	*Eco*RI	GTGGG**AAT**TCTGACCGGCTAC		
*kdg*A	S	*Eco*RI	TAGCGCTGGCC**G**AAT**T**CGCCGAGTCGAG	pSPT19	DH5α
	A	*Bam*HI	ATAGTTGGCCGAGGATCCCACGACTCCG		
*kdg*K	S	*Eco*RI	ACAGGAAGGGGA**ATT**CCGGCAGCAG	pSPT19	DH5α
	A	*Bam*HI	TATGCCTCCTCGGGATCCCTCACTCCGA		
*gap*N[d]				pET15b	BL21(DE3)CP
S. solfataricus					
gad	S	*Nco*I	GCGCGCCATGGCGAGAATCAGAGAAATAGAACCAATAG	pET-24d	BL21(DE3)CP
	A	*Bam*HI	GCGCGGGATCCTCAAACACCATAATTCTTCCAGGTTCCC		
*kdg*A	S	*Nco*I	GCGCGCCATGGCGCCAGAAATCATAACTCCAATCATAACC	pET-24d	JM109(DE3)
	A	*Bam*HI	GCGCGGGATCCCTATTCTTTCAATATTTTAAGCTCTAC		
*kdg*K	S	*Nco*I	GCGCGCCATGGTTGATGTAATAGCTTTGGGAGAGCC	pET-24d	JM109(DE3)
	S	[c]	CTGGGGCTGGTGACGC**A**ATGGCAGGGACATTTGTTTCC		
	A	[c]	GGAAACAAATGTCCCTGCCAT**T**GCGTCACCAGCCCCAG		
	A	*Eco*RI	GCGCGGAATTCTTACGTTTTAAACTCATTTAAAAATC		
*gap*N	S	*Nco*I	GCGCGCCATGGAGAAAACATCAGTGTTG	pET-24d	JM109(DE3)
	A	*Bam*HI	GCGCGGGATCCTTACAAGTATTCCCAAATACCTTTCCC		

[a] Abbreviations: S: sense primer; A: anti-sense primer; RE: introduced restriction site.

[b] The introduced mutations are shown in boldface and the restriction sites are underlined.

[c] For the cloning of *S. solfataricus kdg*K gene, an internal *Nco*I site was disrupted by site-directed mutagenesis, without changing the coding region. The mutated base is indicated in bold.

[d] From reference 44.

Chemicals and Plasmids

All chemicals and enzymes were purchased from Sigma-Aldrich, VWR International or Roche Diagnostics GmbH in analytical grade. ^{14}C- labelled glucose and pyruvate were obtained form Amersham Life Technologies. For heterologous expression the pET vector system (pET-15b, pET-24a, pET-24d; Novagen) was used (Table III).

Heterologous Expression

For expression of the *gdh* (glucose dehydrogenase, AJ621346), *gad* (gluconate dehydratase, AJ621281), *kdg*A (KD(P)G aldolase, AJ621282) and *kdg*K (KDG kinase, AJ 621283) from *T. tenax* and the *gad* (SSO3198), *kdg*A (SSO3197), *kdg*K (SSO3195) and *gap*N (non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPDH), SSO3194) from *S. solfataricus* the genes were cloned into the pET vector system (Novagen) using the restriction sites introduced by PCR mutagenesis (Table III). PCR mutagenesis was performed using Pwo or Taq polymerase (PeQLab, Fermentas) and genomic DNA from *T. tenax* or *S. solfataricus* as template. The sequence of the cloned genes was verified by dideoxy sequencing and expression of the recombinant enzymes in *E. coli* BL21 (DE3), BL21 (DE3) CodonPlus, and JM109 (DE3) was performed following the instructions of the manufacturer (Novagen and Promega, respectively).

Protein Purification

Recombinant *E. coli* cells (1 g wet weight) were suspended in 2 ml of 100 mM HEPES/KOH (pH 7.0, 70°C) containing 7.5 mM dithiothreitol (buffer A) and passed three times through a French pressure cell at 150 MPa. Cell debris and unbroken cells were removed by centrifugation (60,000 x g for 30 min at 4°C). For enrichment the resulting crude extracts were diluted 1:1 with buffer A, and subjected to a heat-precipitation for 30 min at different temperatures. Extracts containing recombinant *T. tenax* and *S. solfataricus* protein were incubated at the following temperatures: *T. tenax* and *S. solfataricus* KD(P)G aldolase and *T. tenax* GAD at 85°C (30 min), *T. tenax* GDH and KDG kinase at 80°C (30 min). Extracts containing recombinant *S. solfataricus* GAPN were incubated at 70°C (20 min), and *S. solfataricus* GAD and KDGK at 65°C (20 min). After heat precipitation the samples were cleared by centrifugation (60,000 x g for 30 min at 4°C). GAPN of *S. solfataricus* was dialyzed overnight against 50 mM HEPES/KOH (pH 7.0, 70°C), 7.5 mM dithiothreitol (2-liter volume, 4°C) and directly used for enzymatic assays. GAD, KD(P)G aldolase, KDG kinase of *T. tenax* and KD(P)G aldolase of *S. solfataricus* were dialyzed overnight (50 mM HEPES/KOH (pH 7.0, 70°C), 7.5 mM dithiothreitol, 300 mM KCl and alternatively for GAD 200 mM KCl) and subjected to gelfiltration on HiLoad 26/60 Superdex 200 prep grade

(Amersham Biosciences) pre equilibrated in the respective buffer. Fractions containing the homogeneous enzyme fraction were pooled and used for enzymatic assays. Calculation of the kinetic parameters (V^{max} and K_m) were performed by iterative curve-fitting (Hanses) using the program Origin (Microcal Software Inc.).

Enzyme Assays

KD(P)G Aldolase and GAD activity were measured at 70°C using the TBA assay as described previously[21] with the following modification. The assay was performed in 100 mM HEPES/KOH (pH 7.0, 70°C), 5 mM pyruvate, and 2 mM D,L-glyceraldehyde or D,L-glyceraldehyde 3-phosphate (GAP) were added as substrates for the KD(P)G aldolase (6 or 12 μg of *T. tenax* and of *S. solfataricus* protein after gelfiltration, respectively; total volume 1 ml). The experiments were performed in the presence of 2 mM GA or GAP and 5 mM pyruvate, since at higher concentrations of GAP (20 mM GAP, 50 mM pyruvate) the observed activity is no more proportional to the amount of enzyme.

GAD activity (30 and 60 μg of *T. tenax* protein after gel filtration; 420/840 μg of *S. solfataricus* protein after heat precipitation; total volume 1 ml) was assayed in the presence of 10 mM gluconate and galactonate. The absorbance was followed at 549 nm ($\varepsilon_{\text{chromophore}}$ = 67.8 x 103 $mM^{-1}cm^{-1}$)[29,53]. Galactonate (10 mM) was prepared from galactonate γ-lactone by incubation in 1 M NaOH for 1 h (4 M stock solution) and subsequent dilution in 50 mM HEPES/KOH (pH 7.0, 70°C) as described[32].

The activity of the *T. tenax* and *S. solfataricus* KDG kinase was determined at 70°C and 60°C, respectively using a continuous assay. The phosphorylation of KDG by ATP was followed by coupling the formation of KDPG to the reduction of NAD^+ via KD(P)G aldolase and GAPN of *T. tenax*[43]. The standard assay was performed in 100 mM HEPES/KOH (pH 7.0, 70°C) in the presence of KDG kinase (1.5 and 3 μg of *T. tenax* protein after gel filtration; 40 and 80 μg of *S. solfataricus* protein after heat precipitation), 2 mM ATP, 2 mM $MgCl_2$, 10 mM NAD^+, KD(P)G aldolase (3 μg of *T. tenax* enzyme) and GAPN (25 μg of *T. tenax* enzyme, protein fraction after heat precipitation). It was assured that the amount of auxiliary enzymes is not rate limiting. The reaction was started by addition of KDG kinase. Enzymatic activities were measured by monitoring the increase in absorption at 340 nm ($\varepsilon_{\text{NADH}}$, 70°C = 5.8 mM^{-1} cm^{-1}).

KDG was synthesized via the KD(P)G aldolase from *T. tenax* as described previously [22]. Briefly, 1 g of D,L- glyceraldehyde and 2.2 g of pyruvate were mixed in 100 ml water containing 6.6 mg of enzyme and the reaction mixture was incubated for 8 h at 70°C. Protein was removed by acetone precipitation and the reaction mix was separated by Dowex 1X8 anion exchange chromatography using a linear 0.0-0.2 M HCl gradient. Fractions containing KDG without glyceraldehydes and pyruvate were identified using the TBA assay, the lactate dehydrogenase assay and TLC. Alternatively, KDG was formed by coupling the reaction to the dehydration of gluconate using gluconate dehydratase from *T. tenax*.

GAPN activity was determined in a continuous assay at 70°C. The standard assay was performed in the presence of 90 mM HEPES/KOH (pH 7.0, 70°C), 160 mM KCl, 3 mM DL-GAP, and 2 mM $NADP^+$[43,44]. The reaction was started by the addition of GAP and the enzyme concentration was 70 μg of protein/ml assay volume (ε_{NADH}, 70°C = 5.8 mM^{-1} cm^{-1}, ε_{NADPH}, 70°C = 5.71 mM^{-1} cm^{-1}).

In vitro assays with crude extracts

Crude extracts of *T. tenax* and *S. solfataricus* cells grown on glucose were prepared as reported for the recombinant *E. coli* cells. After centrifugation the protein solution was dialyzed overnight against 50 mM HEPES/KOH (pH 7.0, 70°C), 7.5 mM dithiothreitol (2-liter volume, 4°C) and directly used for enzymatic assays. Activities in crude extracts (450 or 900 μg of protein; total volume 1 ml) were determined as described for the recombinant proteins.

^{14}C-Labelling Experiments and Thin Layer Chromatography (TLC)

KD(P)G aldolase activity from *T. tenax* and *S. solfataricus* was followed by incubation of dialyzed fractions after heat precipitation (*T. tenax* 16 μg and *S. solfataricus* 11 μg of protein) in 100 mM HEPES/KOH (pH 7.0, 70°C) in the presence of 0.3 μCi [2-^{14}C]pyruvate, 50 mM pyruvate and either 20 mM GA or GAP (total volume 30 μl). A sample was withdrawn before and after incubation at 70°C (30 min) and analyzed by TLC (silica gel G-60 plates without fluorescence indicator (VWR International) developed in butan-1-ol/acetic acid/water (v/v/v = 3/1/1)) and autoradiography (Agfa X-ray 90 films). Intermediates were identified by their R_f-values determined previously [6] and by the formation of KDG and KDPG using the characterized KDPG aldolase (EDA) of *Thermotoga maritima*[34]. The expression plasmid (pTM-eda) was kindly provided by Carol A. Fierke (University of Michigan, Ann Arbor, MI, USA). The enzyme of *T. maritima* was enriched by heat precipitation (30 min 75°C) from the expression host (BL21 (DE3) CodonPlus).

For the *in vitro* reconstruction of the ED pathway, the labelled intermediates were followed after addition of the different ED enzymes (GDH (*T. tenax* 13 μg of protein), GAD (*T. tenax* 11 μg of protein), KD(P)G aldolase (*T. tenax* 23 μg of protein), KDG kinase (*T. tenax* 35 μg of protein). The assay was performed in the presence of 0.3 μCi [U-^{14}C]glucose, 100 mM HEPES/KOH (pH 7.0, 70°C), 10 mM glucose, and 5 mM NADP$^+$. 10 mM ATP and 10 mM Mg^{2+} were added in the presence of KDG kinase. Samples (30 μl volume) were incubated for 10, 30 and 60 min at 70°C and the labelling was followed by TLC as described above.

ACKNOWLEDGEMENTS

This work was supported by the Deutsche Forschungsgemeinschaft. We thank C. A. Fierke for providing the expression plasmid of *T. maritima* KDPG aldolase.

REFERENCES

1 Ronimus, R.S. and Morgan, H.W. (2002) 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

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

3 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**, 231-46

4 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* **377**, 819-822

5 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**, 407-14

6 Budgen, N. and Danson, M.J. (1986) Metabolism of glucose via a modified Entner-Doudoroff pathway in the thermoacidophilic archaebacterium *Thermoplasma acidophilum*, *FEBS Lett* **196**, 207-210

7 Selig, M. and Schönheit, P. (1994) Oxidation of organic compounds to CO2 with sulfur or thiosulfate as electron-acceptor in the anaerobic hyperthermophilic archaea *Thermoproteus tenax* and *Pyrobaculum islandicum* proceeds via the citric acid cycle, *Arch Microbiol* **162**, 286-294

8 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**, 217-32

9 Siebers, B. and Hensel, R. (1993) Glucose catabolism of the hyperthermophilic archaeum *Thermoproteus tenax*, *FEMS Microbiol Lett* **111**, 1-8

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

11 Entner, N. and Doudoroff, M. (1952) Glucose and gluconic acid oxidation of *Pseudomonas saccharophila*, *J Biol Chem* **196**, 853-62

12 Conway, T. (1992) The Entner-Doudoroff pathway: history, physiology and molecular biology, *FEMS Microbiol Rev* **9**, 1-27

13 Szymona, M. and Doudoroff, M. (1960) Carbohydrate metabolism in *Rhodopseudomonas sphreoides*, *J Gen Microbiol* **22**, 167-83

14 Andreesen, J.R. and Gottschalk, G. (1969) The occurrence of a modified Entner-doudoroff pathway in *Clostridium aceticum*, *Arch Mikrobiol* **69**, 160-70

15 Tomlinson, G.A., Koch, T.K. and Hochstein, L.I. (1974) The metabolism of carbohydrates by extremely halophilic bacteria: glucose metabolism via a modified Entner-Doudoroff pathway, *Can J Microbiol* **20**, 1085-1091

16 Elzainy, T.A., Hassan, M.M. and Allam, A.M. (1973) New pathway for nonphosphorylated degradation of gluconate by *Aspergillus niger*, *J Bacteriol* **114**, 457-9

17 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**, 540-8

18 Mukund, S. and Adams, M.W. (1991) The novel tungsten-iron-sulfur protein of the hyperthermophilic archaebacterium, *Pyrococcus furiosus*, is an aldehyde ferredoxin oxidoreductase. Evidence for its participation in a unique glycolytic pathway, *J Biol Chem* **266**, 14208-16

19 Schicho, R.N., Snowden, L.J., Mukund, S., Park, J.B., Adams, M.W.W. and Kelly, R.M. (1993) Influence of tungsten on metabolic patterns in *Pyrococcus furiosus*, a hyperthermophilic archaeon, *Arch Microbiol* **159**, 380-385

20 Grogan, D.W. (1989) Phenotypic characterization of the archaebacterial genus *Sulfolobus*: comparison of five wild-type strains, *J Bacteriol* **171**, 6710-9

21 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**, 563-70

22 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**, 34066-72

23 Zillig, W., Stetter, K.O., Schäfer, W., Janekovic, D., Wunderl, S., Holz, I. and Palm, P. (1981) *Thermoproteales*: a novel type of extremely thermoacidophilic anaerobic archaebacteria isolated from Icelandic solfatares, *Zentbl Bakteriol Hyg 1 Abt Org C* **2**, 205-227

24 Dorr, C., Zaparty, M., Tjaden, B., Brinkmann, H. and Siebers, B. (2003) The hexokinase of the hyperthermophile *Thermoproteus tenax*. ATP-dependent hexokinases and ADP-dependent glucokinases, teo alternatives for glucose phosphorylation in Archaea, *J Biol Chem* **278**, 18744-53

25 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**, 2001-9

26 Babbitt, P.C., Hasson, M.S., Wedekind, J.E., Palmer, D.R., Barrett, W.C., Reed, G.H., Rayment, I., Ringe, D., Kenyon, G.L. and Gerlt, J.A. (1996) The enolase superfamily: a general strategy for enzyme-catalyzed abstraction of the alpha-protons of carboxylic acids, *Biochemistry* **35**, 16489-501

27 Bender, R. and Gottschalk, G. (1973) Purification and properties of D-gluconate dehydratase from *Clostridium pasteurianum*, *Eur J Biochem* **40**, 309-21

28 Bender, R. and Gottschalk, G. (1974) Enzymatic synthesis of 2-keto-3-deoxy-d-glucose from D-gluconate, *Anal Biochem* **61**, 275-9

29 Gottschalk, G. and Bender, R. (1982) D-Gluconate dehydratase from *Clostridium pasteurianum*, *Methods Enzymol* **90 Pt E**, 283-7

30 Kersters, K. and De Ley, J. (1975) D-gluconate dehydratase from *Alcaligenes*, *Methods Enzymol* **42**, 301-4

31 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**, 271-80

32 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**, 133-6

33 Hubbard, B.K., Koch, M., Palmer, D.R., Babbitt, P.C. and Gerlt, J.A. (1998) Evolution of enzymatic activities in the enolase superfamily: characterization of the (D)-glucarate/galactarate catabolic pathway in Escherichia coli, *Biochemistry* **37**, 14369-75

34 Griffiths, J.S., Wymer, N.J., Njolito, E., Niranjanakumari, S., Fierke, C.A. and Toone, E.J. (2002) Cloning, isolation and characterization of the *Thermotoga maritima* KDPG aldolase, *Bioorg Med Chem* **10**, 545-50

35 Babbitt, P.C. and Gerlt, J.A. (1997) Understanding enzyme superfamilies. Chemistry As the fundamental determinant in the evolution of new catalytic activities, *J Biol Chem* **272**, 30591-4

36 Theodossis, A., Walden, H., Westwick, E.J., Connaris, H., Lamble, H.J., Hough, D.W., Danson, M.J. and Taylor, G.L. (2004) The structural basis for substrate promiscuity in 2-keto-3-deoxygluconate aldolase from the Entner-Doudoroff pathway in *Sulfolobus solfataricus*, *J Biol Chem* **279**, 43886-92

37 Peekhaus, N. and Conway, T. (1998) What's for dinner? Entner-Doudoroff metabolism in *Escherichia coli*, *J Bacteriol* **180**, 3495-502

38 Bork, P., Sander, C. and Valencia, A. (1993) Convergent evolution of similar enzymatic function on different protein folds: the hexokinase, ribokinase, and galactokinase families of sugar kinases, *Protein Sci* **2**, 31-40

39 Ito, S., Fushinobu, S., Yoshioka, I., Koga, S., Matsuzawa, H. and Wakagi, T. (2001) Structural basis for the ADP-specificity of a novel glucokinase from a hyperthermophilic archaeon, *Structure (Camb)* **9**, 205-14

40 Cynkin, M.A. and Ashwell, G. (1960) Uronic acid metabolism in bacteria. IV. Purification and properties of 2-keto-3-deoxy-D-gluconokinase in *Escherichia coli*, *J Biol Chem* **235**, 1576-9

41 Hugouvieux-Cotte-Pattat, N., Nasser, W. and Robert-Baudouy, J. (1994) Molecular characterization of the *Erwinia chrysanthemi kdg*K gene involved in pectin degradation, *J Bacteriol* **176**, 2386-92

42 Shulami, S., Gat, O., Sonenshein, A.L. and Shoham, Y. (1999) The glucuronic acid utilization gene cluster from *Bacillus stearothermophilus* T-6, *J Bacteriol* **181**, 3695-704

43 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**, 6149-56

44 Lorentzen, E., Hensel, R., Knura, T., Ahmed, H. and Pohl, E. (2004) Structural Basis of allosteric regulation and substrate specificity of the non-phosphorylating glyceraldehyde 3-Phosphate dehydrogenase from Thermoproteus tenax, *J Mol Biol* **341**, 815-28

45 Pohl, E., Brunner, N., Wilmanns, M. and Hensel, R. (2002) The crystal structure of the allosteric non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase from the hyperthermophilic archaeum *Thermoproteus tenax*, *J Biol Chem* **277**, 19938-45

46 Ahmed, H., Tjaden, B., Hensel, R. and Siebers, B. (2004) Embden-Meyerhof-Parnas and Entner-Doudoroff pathways in *Thermoproteus tenax*: metabolic parallelism or specific adaptation? *Biochem Soc Trans* **32**, 303-4

47 Siebers, B., Klenk, H.P. and Hensel, R. (1998) PPi-dependent phosphofructokinase from *Thermoproteus tenax*, an archaeal descendant of an ancient line in phosphofructokinase evolution, *J Bacteriol* **180**, 2137-43

48 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., Erauso, G., Fletcher, C., Gordon, P.M., Heikamp-de Jong, I., Jeffries, A.C., Kozera, C.J., Medina, N., Peng, X., Thi-Ngoc, H.P., Redder, P., Schenk, M.E., Theriault, C., Tolstrup, N., Charlebois, R.L., Doolittle, W.F., Duguet, M., Gaasterland, T., Garrett, R.A., Ragan, M.A., Sensen, C.W. and Van der Oost, J. (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2, *Proc Natl Acad Sci U S A* **98**, 7835-40

49 Zillig, W., Stetter, K.O., Wunderl, S., Priess, H. and Scholz, J. (1980) The *Sulfolobus*-"Caldariella" group: taxonomy on the basis of the structure of DNA-dependent RNA polymerases, *Arch Microbiol* **125**, 259-269

50 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**, 29537-49

51 Brunner, N.A., Siebers, B. and Hensel, R. (2001) Role of two different glyceraldehyde-3-phosphate dehydrogenases in controlling the reversible Embden-Meyerhof-Parnas pathway in *Thermoproteus tenax*: regulation on protein and transcript level, *Extremophiles* **5**, 101-9

52 Sambrook, J., Fritsch, E.F. and Maniatis, T. (1989) *Molecular Cloning - A Laboratory Manual*, Cold Spring Harbor

53 Skoza, L. and Mohos, S. (1976) Stable thiobarbituric acid chromophore with dimethyl sulphoxide. Application to sialic acid assay in analytical de-O-acetylation, *Biochem J* **159**, 457-62

Chapter 9

Summary and concluding remarks

In the mid-seventies, Carl Woese and co-workers discovered a class of cellular organisms, referred to as ‘archaebacteria’, which could neither be classified as Bacteria nor Eukaryotes. Later, these organisms were proposed to constitute a separate, third domain of life: the domain of the Archaea. Early pioneering research on Archaea validated the classification of the Archaea as being a separate prokaryotic domain, and revealed many surprising and unique features of their mysterious biology. At the beginning of the ‘genomic era’, the first completely sequenced genomes, especially the archaeal genomes, revealed mysterious genomescapes, containing many genes for which no, or only a general function could be assigned. Clearly, there was a need for methods that enabled for rational-based function prediction, helping experimentalists to efficiently design their experimental setup. In order to meet this need, soon the field of ‘functional genomics’ emerged. The *rationale* of functional genomics is to increase insight in function of a particular biological system by comparing and integrating the ever increasing amount of genomics data.

In general, the main theme in this thesis is the function prediction and verification of novel archaeal systems, using both a combined functional and comparative genomics approach: by comparing and integrating the available archaeal genomics data, we have revealed novel and interesting aspects of the biology and evolution of the Archaea. The first scientific chapter of this thesis, **Chapter 2**, gives an up-to-date overview about how the conceptually different comparative genomics tools have contributed in this.

In **Chapter 3** is described how, for the first time, a comparative genomics approach was used to gain insight in the evolution of genome content, and how this information can be used to refine the prediction of protein function. A comprehensive comparison of the sequenced genomes of three closely related archaeal species (*Pyrococcus furiosus*, *P. abyssi* and *P. horikoshii*) revealed insight in how these genomes have evolved since their radiation from their last common ancestor. The evolution of these genomes can be described in terms of gain and loss of genes and the presented results indicate that this process is modular: genes that are functionally linked tend to be gained or lost together during genome evolution. For example, *P. furiosus* has enhanced its carbohydrate utilizing potential by acquiring a considerable amount of genes encoding for carbohydrate metabolizing enzymes, reflecting its adaptation towards a more specialized heterotrophic lifestyle. Interestingly, the process of gene loss appears to be more modular than the gain of genes. Exceptions to this pattern of modularity hint at relatively distant functional interactions between the encoded proteins, enabling the refinement of protein function prediction.

Chapter 4 describes how a comparative genomics approach can be used to identify novel functions. By comparing the gene context of multiple archaeal genomes, a conserved gene cluster was detected that potentially encoded a novel type of heavy metal resistance system. The conserved cluster comprises genes that encode a P-type cation transporting ATPase (*cop*A), a novel type of metallochaperone (*cop*M) and a predicted transcriptional regulator (*cop*T). The protein encoded by the latter gene is anticipated to be involved in the regulation of the gene cluster. Remarkably, the proteins that are encoded by the gene cluster share a conserved cystein motif that is predicted to be involved in the binding of metal ions. Since the novel motif is anticipated to be involved in the trafficking, resistance and sensing of heavy metals, it was named the 'TRASH domain'.

Chapter 5 describes how the *in silico* prediction of the archaeal *cop* gene cluster (Chapter 4) is functionally investigated using an experimental approach. The study concerns the molecular characterization of the *cop* gene cluster of the thermoacidophilic crenarchaeon *Sulfolobus solfataricus*. The transcription start sites of *cop*T and *cop*MA have been mapped and the polycistronic *cop*MA transcript was found to accumulate in response to growth-inhibiting copper concentrations, whereas *cop*T transcript abundance appeared to be unaffected by copper supplementation. Furthermore, electrophoretic mobility shift assays (EMSAs) revealed that CopT binds to the *cop*MA promoter, forming at least five different complexes. DnaseI footprinting studies revealed the presence of multiple CopT binding sites, both upstream and downstream of the predicted TATA-BRE site. These results describe the first copper-responsive operon in archaea, a new family of archaeal DNA-binding proteins, and support a prominent role of the original TRASH domain in archaeal copper response.

In addition to comparing genomes, the comparative study of protein structures can yield valuable information regarding protein function. An example that illustrates this is described in **Chapter 6**. Here, the starting point was the resolved crystal structure of LrpA, a transcriptional regulator of the hyperthermophilic euryarchaeon *P. furiosus*. By using a combination of sequence profile searching and structural protein analysis we detected a novel type of small molecule binding domain, which potentially functions as an allosteric regulatory switch in prokaryotes. The domain, designated RAM (after regulatory domain of amino acid metabolism), has been found as a fusion with the DNA-binding domain of Lrp-like transcription regulators and with the catalytic domain of some metabolic enzymes. In addition, it is also found as a stand-alone module. Structural analysis of the RAM domain of LrpA reveals a βαββαβ-fold that is strikingly similar to that of the recently described ACT domain, a ubiquitous allosteric regulatory domain of many metabolic enzymes that is present in all three domains of life. However, structural alignment and re-evaluation of previous mutagenesis data suggest that the effector-binding sites of both modules are significantly different. Although it cannot be ruled out that the RAM and ACT domains originate from a common ancestor, our observations suggest that their ligand-binding sites have evolved independently. However, both domains appear to play analogous roles in controlling key steps in amino acid metabolism at the level of gene expression as well as enzyme activity.

In **Chapter 7**, a bioinformatics approach is used to identify a missing link in archaeal central carbon metabolism, the archaeal variant of phospho*enol*pyruvate carboxylase (atPEPC). Despite the fact that PEPC activity has been measured and in some cases even purified from some Archaea, the gene responsible for this activity had never been elucidated. By using sensitive sequence comparison methods, we were able to detect a highly conserved, uncharacterized archaeal gene family that was distantly related to the catalytic core of the canonical PEPC, present in Bacteria and Eukarya. Subsequent functional analysis of the representative of this gene family from the hyperthermophilic acidophile *S. solfataricus* confirmed that the encoded protein indeed displayed highly thermostable PEPC activity. The newly identified archaeal PEPC, with its distinct properties, constitutes yet another example of the versatility of the enzymes of the central carbon metabolic pathways in the archaeal domain.

Finally, in **Chapter 8**, we investigated the modified Entner-Doudoroff (ED) pathways of hyperthermophilic archaea, which, until this research, were generally believed to proceed via a non-phosphorylative variant. A comparative genomics analysis revealed the presence of an ED gene cluster in multiple archaeal genomes, among which the genomes of *Thermoproteus tenax* and *S. solfataricus*. In addition to the genes that constitute the core supportive of a non-phosphorylative ED variant, also genes encoding a potential 2-keto-3-deoxygluconate kinase (*kdg*K) and non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (*gap*N) were identified. These findings prompted us to investigate a possible operation of a semi-phosphorylative ED pathway in these hyperthermophiles. Indeed, transcription analysis, enzymatic studies as well as *in vitro* reconstruction experiments indicate the operation of such a semi-phosphorylative ED pathway in *T. tenax* and *S. solfataricus*. This study is an example of how a genome-based comparative approach can provide new information, prompting for a re-evaluation of past knowledge.

The work presented in this thesis demonstrates that comparative genome analysis is extremely useful for gaining insight in the lifestyle, biology and evolution of the genetically less tractable Archaea. In addition, this work shows the efficacy of combining *in silico* analysis with experimental work. In the beginning of this project, in September 2000, only a handful of archaeal genomes, mostly from thermophilic euryarchaeotes, were available for analysis. However, despite the limited number of genomes, their restricted diversity as well as their poor annotation, it was possible to extract valuable information from them, as was demonstrated in Chapter 2. In the years that followed, the collection of archaeal genomes matured into a phylogenetically and physiologically diverse subset. These genomes have proven to be a goldmine for archaeal researchers and they greatly enhanced the sensitivity of the comparative analyses, revealing several interesting features and novel archaeal systems, some of which are described in this thesis (Chapters 4, 5, 7 and 8).

The field of archaeal research is very much alive, and the number of researchers working in this field is growing steadily year by year. For this growing archaeal community, there are still many challenges remaining in the near future, since many questions about the biology of the Archaea remain unanswered. For example, we are only beginning to understand their diverse metabolic repertoire and the mechanisms that underlie basal processes as replication, transcription and translation. However, not much is known about the regulatory networks that are involved in the global regulation and the fine tuning of these processes.

Currently, one of the major issues is the lack of efficient archaeal model organisms[1]. The development of efficient genetic systems will be a key-objective in order to perform proper molecular genetics on archaeal species (e.g. making knock-outs and homologous recombination). Other challenges are to be expected from the isolation of new archaeal species. The recent discovery of a totally new archaeal clade, the *Nanoarchaea*, proves that there are still some pleasant surprises out there. What to think of the mysterious Korarchaea[2], another deeply branching archaeal clade, detected by phylogenetic analysis of ribosomal RNA sequences obtained from uncultivated organisms of a hot spring? Recently, Joint Genome Insitue (DOE-JGI, California, USA) has initiated an environmental shotgun sequencing project of Obsidian Pool, a hotspring in Yellowstone National Park, USA that is known to contain a diverse archaeal community, including some korarchaeal species. These types of projects are likely to uncover some of the secrets about the diversity and biology of the organisms that reside in these archaea-dominated thermal environments.

A GLIMPSE IN THE FUTURE...

Recent developments in the field of functional genomics have resulted in more powerful techniques, enabling for large-scale and high-throughput experimental design. The availability of these powerful techniques allow for the development of new research strategies that, by integration of different genomics disciplines, are able to study complete biological systems. This cross-disciplinary, holistic approach is referred to as 'Systems Biology' (Figure 9.1). The key disciplines of systems biology are metabolomics (analysis of metabolites, metabolic fluxes and pathways), proteomics (analysis of expressed proteome), structural genomics (analysis of structure-function relations) and high through-put experimental applications (e.g. genome-wide transcription analysis or protein-protein interaction maps). Last but not least, computational biology will play a crucial role in analyzing and integrating the data that is generated in these large-scale experiments.

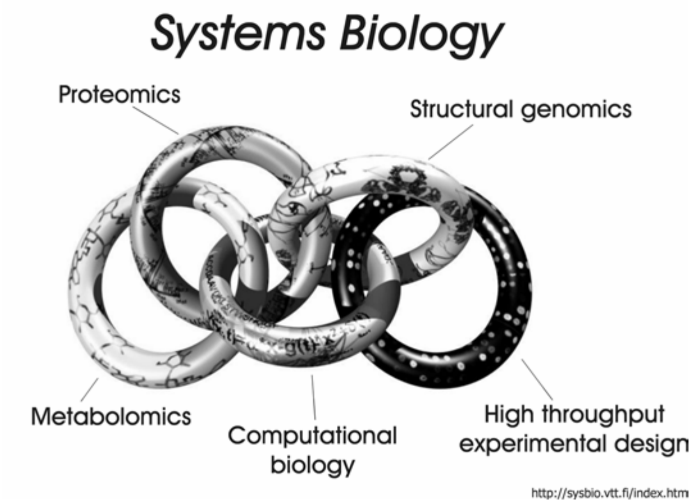


Figure 9.1 Schematical impression of the interdependence of the disciplines that are part of a Systems Biology approach.

The set-up of archaeal systems biology initiatives will accelerate and revolutionize the research on the intriguing biology of the Archaea. The initial results of such an initiative[3-7], which has focused on the biology of the extreme halophilic archaeon *Halobacterium* sp. NRC-1, are more than promising and prove the efficacy of such an approach.

REFERENCES

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

2 Barns, S.M., Delwiche, C.F., Palmer, J.D. and Pace, N.R. (1996) Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences, *Proc Natl Acad Sci U S A* **93**, 9188-93

3 Baliga, N.S., Pan, M., Goo, Y.A., Yi, E.C., Goodlett, D.R., Dimitrov, K., Shannon, P., Aebersold, R., Ng, W.V. and Hood, L. (2002) Coordinate regulation of energy transduction modules in Halobacterium sp. analyzed by a global systems approach, *Proc Natl Acad Sci U S A* **99**, 14913-8

4 Baliga, N.S., Bjork, S.J., Bonneau, R., Pan, M., Iloanusi, C., Kottemann, M.C., Hood, L. and DiRuggiero, J. (2004) Systems level insights into the stress response to UV radiation in the halophilic archaeon Halobacterium NRC-1, *Genome Res* **14**, 1025-35

5 Baliga, N.S., Bonneau, R., Facciotti, M.T., Pan, M., Glusman, G., Deutsch, E.W., Shannon, P., Chiu, Y., Weng, R.S., Gan, R.R., Hung, P., Date, S.V., Marcotte, E., Hood, L. and Ng, W.V. (2004) Genome sequence of Haloarcula marismortui: a halophilic archaeon from the Dead Sea, *Genome Res* **14**, 2221-34

6 Bonneau, R., Baliga, N.S., Deutsch, E.W., Shannon, P. and Hood, L. (2004) Comprehensive de novo structure prediction in a systems-biology context for the archaea Halobacterium sp. NRC-1, *Genome Biol* **5**, R52

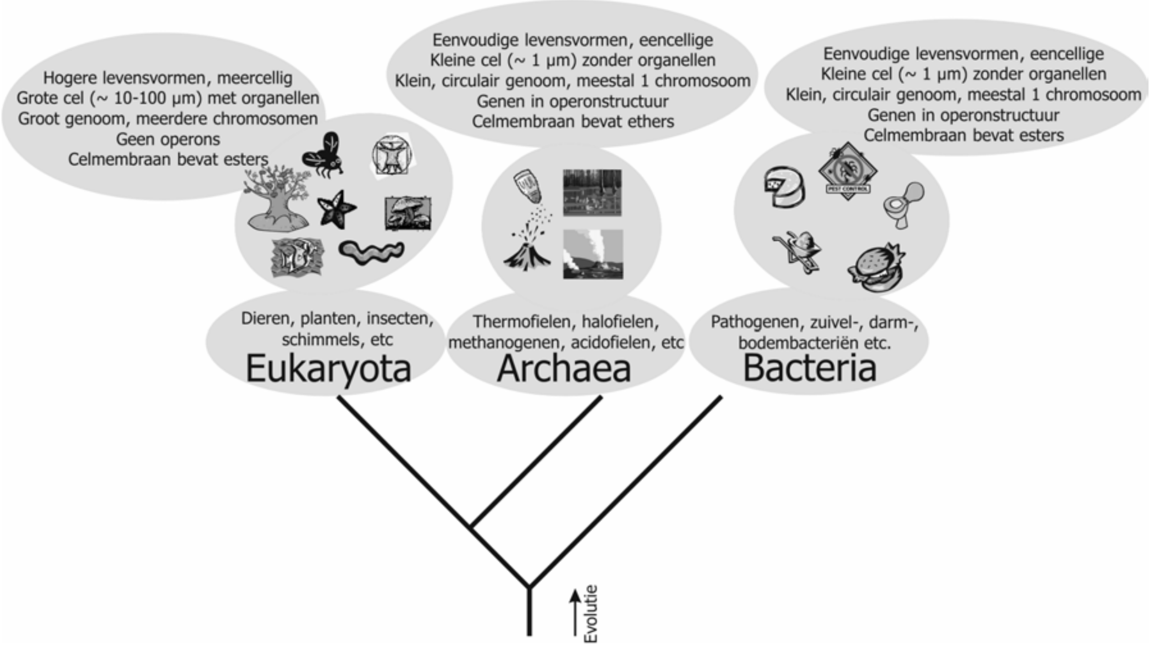
7 Goo, Y.A., Yi, E.C., Baliga, N.S., Tao, W.A., Pan, M., Aebersold, R., Goodlett, D.R., Hood, L. and Ng, W.V. (2003) Proteomic Analysis of an Extreme Halophilic Archaeon, Halobacterium sp. NRC-1, *Mol Cell Proteomics* **2**, 506-24

Chapter10

Nederlandse samenvatting

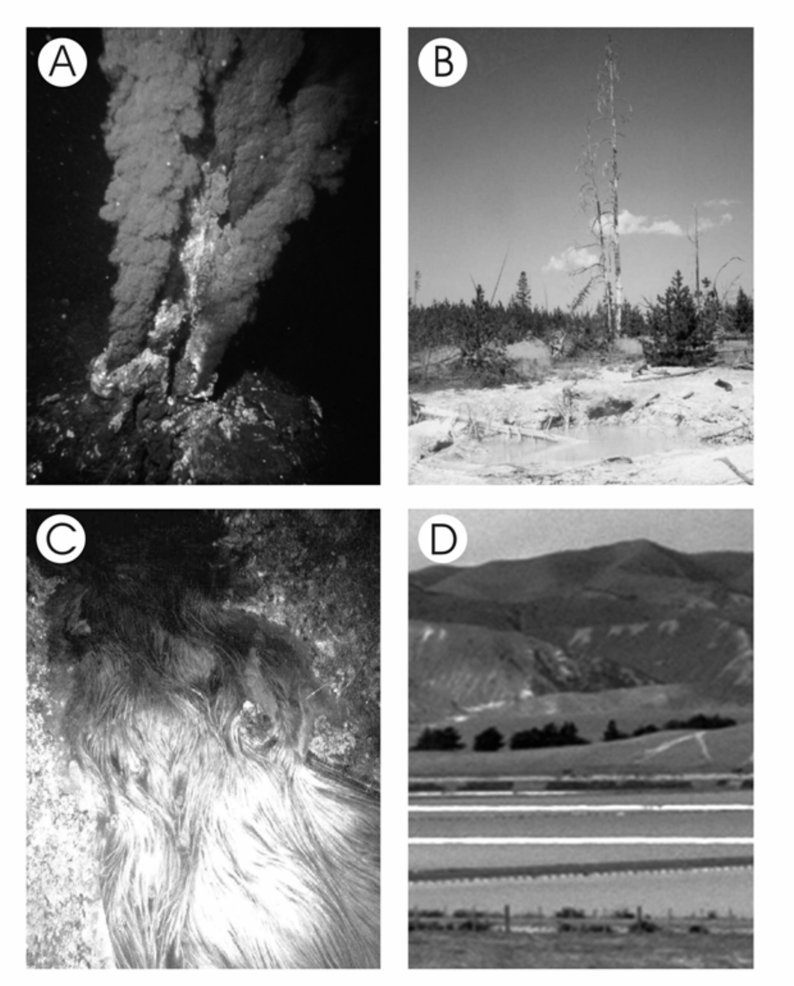
Een nieuwe vorm van leven?

Halverwege de jaren zeventig ontdekte de evolutiebioloog Carl Woese een bepaald type micro-organismen, ‘archaebacteriën’, die hij niet kon thuisbrengen in de destijds gangbare klassificatie van de bacteriën (eenvoudige eencelligen) en de eukaryoten (complexere levensvormen). Enkele jaren later stelde Woese voor om op basis van vergelijkend DNA onderzoek deze ‘vreemde’ micro-organismen te klassificeren als een apart derde domein van het leven op aarde, de Archaea, naast de domeinen van de Eukarya en de Bacteria (Figuur 10.1). Het spreekt voor zich dat dit voorstel destijds veel stof (tot nadenken) deed opwaaien… In 2003 ontving Woese de befaamde Craford prijs (‘de Nobelprijs voor de biologie’) voor zijn ontdekking van de Archaea.


Hogere levensvormen, meercellig
Grote cel (~ 10-100 μm) met organellen
Groot genoom, meerdere chromosomen
Geen operons
Celmembraan bevat esters
Eenvoudige levensvormen, eencellige
Kleine cel (~ 1 μm) zonder organellen
Klein, circulair genoom, meestal 1 chromosoom
Genen in operonstructuur
Celmembraan bevat ethers
Eenvoudige levensvormen, eencellige
Kleine cel (~ 1 μm) zonder organellen
Klein, circulair genoom, meestal 1 chromosoom
Genen in operonstructuur
Celmembraan bevat esters
PEST CONTROL
Dieren, planten, insecten, schimmels, etc
Eukaryota
Thermofielen, halofielen, methanogenen, acidofielen, etc
Archaea
Pathogenen, zuivel-, darm-, bodembacteriën etc.
Bacteria
Evolutie

Figuur 10.1 De drie domeinen van het leven zoals die in 1990 door Carlo Woese werden voorgesteld: De Eukayota, waartoe alle hogere levensvormen behoren, maar ook ééncelligen als gist en het pantoffeldiertje. Geheel rechts zien we de Bacteria, ééncellige organismen, voor het eerst waargenomen door Antonie van Leeuwenhoek in het begin van de achttiende eeuw. Binnen het domein van de bacteriën bestaat een grote diversiteit: sommigen worden gebruikt voor de bereiding van voedsel (bijv. melkzuurbacteriën), andere zijn verantwoordelijk voor wereldwijde epidemieën van de pest. In het midden zien we de tak van de Archaea, waartoe de veelal extremofielen behoren, maar niet uitsluitend. In sommige aspekten zijn de Archaea meer verwant aan de Eukaryota, in anderen lijken ze weer meer op de Bacteria.

Extremofielen, of toch niet...



Figuur 10.2 (NB: Voor een KLEUREN versie van deze figuur zie APPENDIX I) Extreme milieu's waar sommige archaea voorkomen: **(a)** *Mid-atlantic ridge, Atlantische oceaan* – Deze donkere pluimen van hete thermale vloeistof zijn het gevolg van onderaardse vulkanische activiteit en rijk aan opgeloste metalen. Als deze opgeloste metalen in contact komen met het koude zeewater, wordt een zwarte neerslag gevormd. Deze neerslag zinkt naar de zeebodem en creëert zo een schoorsteenachtige structuur, de 'black smoker'. De thermale vloeistoffen zijn rijk aan nutriënten, wat bepaalde organismen in staat stelt om in deze extreme hitte te leven, de (hyper)*thermofielen* [Foto: Peter Rona, NOAA]. **(b)** *Ragged Hills, Yellowstone National Park, USA* – Dit thermale gebied in Yellowstone NP is nog relatief jong en is ontstaan als gevolg van recente thermale activiteit. Door de stijging van temperatuur en zuurtegraad sterven de bomen en planten af en gaan hotsprings het landschap domineren. Deze hotsprings zijn vaak kokend heet en enorm zuur (pH < 3). In dit type hotsprings leven acidofielen als *Sulfolobus* en *Thermoplasma* [Foto: Thijs Ettema]. **(c)** *Iron Mountain, California, USA* – Waterstroom afkomstig uit een oude ijzer/koper mijn, die zich het beste laat vergelijken met accuzuur: extreem zuur (pH tussen 1.3 and 2.2) en extreem hoge metaal concentraties (tot 120 gram per liter). In deze extreme condities leven metallofielen als Ferroplasma, die energie winnen door metaalsulfides te oxideren [Foto: Katrina Edwards, WHOI]. **(d)** *Blenheim, Nieuw Zeeland* – Het zoutmeer op deze foto wordt gebruikt voor commerciële zoutwinning door middel van verdamping van zeewater, waardoor het zout zijn verzadigingspunt (30%) bereikt en uitkristalliseert. Deze extreem zoute meren, waar de zout concentratie 10 keer zo hoog is als in de zee, zijn vaak paars of roze van kleur. Deze kleur wordt veroorzaakt door micro-organismen die in staat zijn deze extreme zoute condities te weerstaan, de halofielen. De meest extreme halofielen behoren tot de archaea [Foto: Thijs Ettema].

Het eerste pionieronderzoek aan soorten behorend tot deze mysterieuze archaea liet zien dat we hier inderdaad te maken hadden met bijzondere organismen, met soms verrassende kenmerken. Sommige soorten werden geïsoleerd op plaatsen waar leven voorheen onmogelijk geacht werd. Zo werden ze in grote getalen aangetroffen in de nabijheid van onderzeese aktieve vulkanen (‘black smokers’), in vaak extreem zure, soms kokende modderpoelen (hotsprings), in zoutmeren verzadigd met zout, in verlaten mijnschachten met extreme -voor de mens toxische- metaalconcentraties (vergelijkbaar met accuzuur), enzovoorts. Hierdoor hebben archaea ookwel het predikaat ‘extremofielen’ (=houdend van extremen) opgeplakt gekregen (Figuur 10.2). Tegenwoordig weten we echter dat archaea zich qua leefomgeving niet beperken tot deze extreme omstandigheden, maar dat zij overal ter wereld in de meest diverse milieu’s voorkomen, zelfs in onze eigen darmen! Vanuit meerde oogpunten zijn archaea vervolgens gebombardeerd tot interessante onderzoeksvoorwerpen. Evolutionair gezien zijn archaea interessant omdat ze in verband gebracht worden met het onstaan van het eerste leven op aarde. Vanuit biotechnologisch oogmerk zijn archaea ook interessant: de extreme omstandigheden waartegen sommige soorten bestand zijn vereisen stabiele bouwstenen (eiwitten, metabolieten), die van toepassing kunnen zijn in bijvoorbeeld de bio- en nanotechnologie.

EXTREMOFIELEN — Micro-organismen die leven in extreme omstandigheden, zoals extreme hitte (‘hyperthermofielen’), zoutconcentratie (‘halofielen’) of zuurtegraad (‘acidofielen’). De meeste extremofielen behoren tot het domein van de Archaea.

De unieke biologie van de archaea

Sinds de ontdekking van de archaea is erg veel onderzoek verricht met als doel om de (moleculaire) biologie van deze organismen beter in kaart te brengen. Initieel molecular biologisch onderzoek wees uit dat de systemen die betrokken zijn bij het kopiëren, aflezen en vertalen van het genetisch materiaal (zie ook Box 1, “Het centrale dogma van de moleculaire biologie”) in archaea meer verwant zijn aan die van eukaryoten dan aan die van bacteriën. Dit in tegenstelling tot hun bacteriële uiterlijk (ééncellige levenswijze, het ontbreken van celorganellen) en de organisatie van hun genetisch materiaal op een kompakt circulair genoom waarop de genen geklusterd liggen in operons (zie ook Figuur 10.3).

Ook de fysiologie van de archaea bleek vol verrassingen te zitten. Al reeds lange tijd

GENOOM — Term die wordt gebruikt om het totale erfelijke materiaal van een organisme aan te duiden. In het geval van een micro-organisme omvat het genoom meestal 1 circulair chromosoom; het menselijke genoom omvat 46 chromosomen.

OPERON — Klustering op het genoom van genen met een soortgelijke functie, met als doel om de regulatie van de expressie van deze genen te vereenvoudigen.

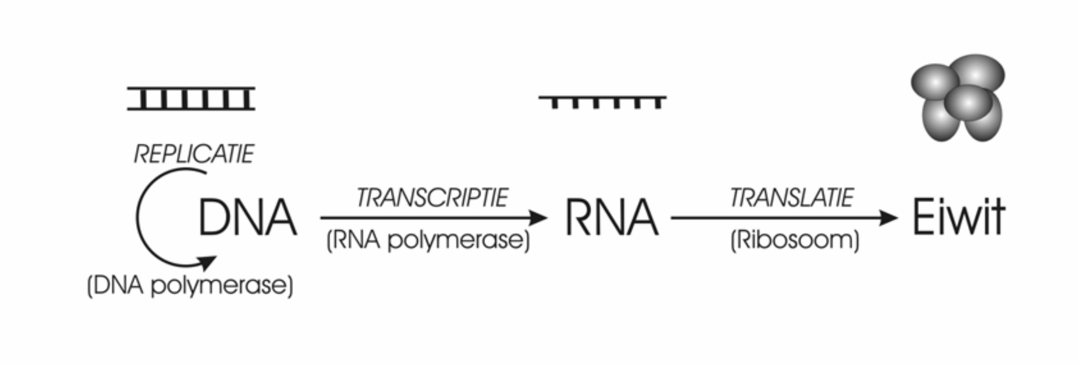
was bekend dat het proces van methanogenese (vorming van methaan) exclusief door bepaalde archaea ('methanogenen') kan worden gedaan.

Box1: Het *centrale dogma* van de moleculaire biologie

Het 'centrale dogma van de moleculaire biologie werd voor het eerst uiteen gezet door Francis Crick in 1958, en in een Nature publicatie uit 1970 geherformuleerd. De officiële definitie luidt:

"Het centrale dogma van de moleculaire biologie behelst het gedetailleerd, stap voor stap overbrengen van sequentiële informatie."

Meestal wordt deze stelling echter versimpeld geïnterpreteerd als volgt:



In feite duidt de stelling dus op de overdracht van informatie die opgeslagen ligt op het DNA, wat zichzelf in stand houdt via replicatie. Op het DNA liggen genen gecodeerd die via het transcriptieproces overgeschreven worden in RNA. Dit RNA wordt vervolgens vertaald naar de aminozuurvolgorde die hierop gecodeerd ligt. Volgens Crick's originele definitie kan, als de informatie eenmaal het stadium van eiwit heeft bereikt, deze informatie niet meer terugvloeien vaar RNA of DNA. De overdracht van informatie is dus unidirectioneel (al zijn hier uitzonderingen op in de vorm van retro- en RNA virussen).

Archaea zijn dus eigenlijk direct verantwoordelijk voor het gas wat we dagelijks gebruiken om voesel te bereiden! Fysiologisch gezien zijn archaea ook in andere aspekten unieke en inventieve organismen gebleken, die het vaak 'net even anders' doen als gebruikelijk. Zo zijn hun metabole routes doorspekt met enzymen die geenszins lijken op de gangbare enzymen zoals we die vinden in bacteriën en eukaryoten.

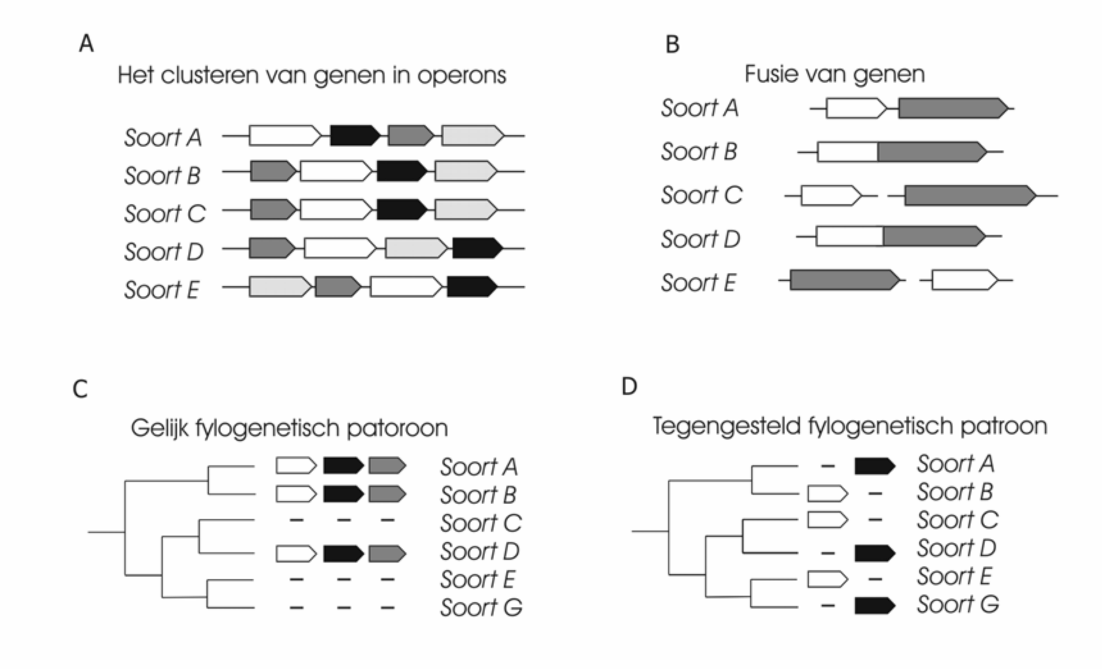
Archaea in en na het 'genoom tijdperk'

Halverwege de jaren negentig zijn moleculair-biologische technologieën zover doorontwikkeld dat het mogelijk werd om de DNA sequentie van het totale erfelijke materiaal van een organisme, het genoom, op te helderen.

GENOOMSEQUENTIE — Volgorde van alle nucleotide basen (A, C, T en G) van het complete genoom van een organisme. Ter vergelijking, een gemiddeld genoom van een bacterie of archaeon bestaat ongeveer uit 2-3 miljoen baseparen, terwijl het genoom van de mens meer dan 3,2 miljard baseparen (!) telt

In deze periode, ookwel het het 'genoom tijdperk' genoemd, werden in een hoog tempo genoomsequenties van een aantal organismen opgehelderd, zo ook van het archaeon *Methanocaldococcus jannaschii*, een thermofiel methanogeen. Tijdens de analyse van deze eerste archaeale genoomsequentie werd al snel duidelijk dat het unieke karakter van de archaea ook weerspiegeld werd in hun genoomsequentie; sterker nog, aan het overgrote deel van de genen die gecodeerd lagen op dit 'mysterieuze genoom landschap' kon geen functie gekoppeld worden. Klaarblijkelijk lag er nog een grote taak weggelegd voor onderzoekers van de archaeale

Box 2: Extra informatie uit genoomdata



Figuur 10.3 Er zijn enkele conceptueel verschillende methoden ontwikkeld die gebruik maken van genoom data om de voorspelling van functies van genen en eiwitten te verbeteren. Enkele van deze methoden maken gebruik van het clusteren van genen die functioneel verwant zijn, bijvoorbeeld doordat ze deel uitmaken van hetzelfde eiwit complex, of metabole route. Het clusteren van genen gebeurt meestal in operons **(a)**, maar soms ook directer, door fusie van 2 of meer genen **(b)**. Als een cluster of een fusie van genen in meerdere, verschillende soorten voorkomt, dan is dit een betrouwbare indicatie dat de eiwitten die door deze genen gecodeerd worden iets met elkaar te maken hebben. Genen die betrokken bij hetzelfde proces ('functioneel gelinkt zijn'), komen met een grote waarschijnlijkheid ook voor op eenzelfde genoom, omdat bij het ontbreken van één of meer van deze genen, dit proces niet optimaal zal verlopen of zelfs stagneren. Door het aan- of afwezig zijn van genen in kaart te brengen voor de genomen van verschillende organismen worden zogenaamde fylogenetische patronen verkregen **(c)**. Hebben 2 genen hetzelfde fylogenetische patroon, dan is het waarschijnlijk dat ze betrokken zijn bij hetzelfde proces. Echter, als 2 genen een tegenovergesteld fylogenetisch patroon vertonen (**'**anti-correlatie', **d)**, dan zou dit erop kunnen duiden dat ze exact dezelfde functie hebben: soms coderen 2 verschillende, niet homologe genen voor dezelfde functie, bijvoorbeeld een enzym met dezelfde activiteit. Het is dan voordeliger voor een organisme om òf het ene gen òf het andere gen te hebben. In de loop van de evolutie komt het soms voor dat het ene gen door het andere wordt vervangen via een proces wat 'non-homologous gene displacement' wordt genoemd.

gemeenschap om de vele vragen die het ophelderen van deze genoomsequentie opriep, te beantwoorden.

Naarmate er meer genoomsequenties beschikbaar kwamen, groeide het besef dat er in deze genomen ‘extra informatie’ verborgen zit, die van pas kan komen om funkties van genen te voorspellen. In box 2 (Figuur 10.3) staat uitgelegd wat voor ‘extra informatie’ er verscholen ligt in deze genoomsequenties en hoe hier functionele informatie uit kan worden afgeleid. Het vakgebied dat bezighoudt met het voorspellen van functies van genen en eiwitten aan de hand van genoom informatie heet ‘functional genomics’. Door meerdere genomen tegelijkertijd te bestuderen, en deze te vergelijken, kunnen voorspellingen betrouwbaarder maar ook nauwkeuriger worden. Het vakgebied dat bezighoudt met het vergelijken van genomen wordt ookwel ‘comparative genomics’ genoemd (vergelijkende genoomleer).

In dit proefschrift…

… is getracht om aan de hand van een aantal beschikbare archaeale genoomsequenties meer inzicht te krijgen in bepaalde aspekten van de biologie en evolutie van de archaea. Het analyseren en vergelijken van deze genoomsequenties heeft geresulteerd in een aantal interessante bevindingen, waarvan er enkele met behulp van laboratorium experimenten verder bestudeerd zijn. De resultaten van deze analyses en experimenten staan beschreven in dit proefschrift.

In ***hoofdstuk 1*** van dit proefschrift wordt het doel en de algehele lijn van het proefschrift uiteengezet, gevolgd door een overzicht (review) van de recente ontwikkelingen in het veld van de comparative en functional genomics met betrekking tot de archaea (***hoofdstuk 2***).

Jongleren met genen

In ***hoofdstuk 3*** is getracht om inzicht te krijgen in de genoom evolutie van 3 verwante hyperthermofiele archaeale soorten, *Pyrococcus furiosus*, *P. abyssi* and *P. horikoshii*. Uit een analyse van de genoomsequenties van deze 3 soorten bleek dat, ondanks de grote verwantschap tussen deze soorten, er toch aanzienlijke verschillen te zien zijn in hun genomen. Het analyseren van deze verschillen in termen van het verkrijgen

GEN — Stuk DNA dat codeert voor een erfelijke eigenschap, meestal een structureel eiwit of enzym.

nieuwe en het verliezen van oude genen heeft geresulteerd in inzicht hoe deze soorten zijn geëvolueerd na de afsplitsing van hun laatste gemeenschappelijke voorouder. Uit deze analyse bleek dat genen die coderen voor eiwitten die een soortgelijke functie hebben, vaak tegelijkertijd verloren gaan of juist verkregen worden. Zo bleek onder andere dat *P. furiosus* sinds de afsplitsing van de gemeenschappelijke *Pyrococcus* voorouder een aantal genen te hebben verkregen die coderen voor suikerafbrekende enzymen. Het vergroten van zijn suikerafbrekend potentieel kan gezien worden als een aanpassing aan de leefomstandigheden van dit organisme. Verder wordt in dit hoofdstuk beschreven hoe patronen van verlies en aanwinst van genen gebruikt kunnen worden om de voorspellingen van functies eiwitten te verbeteren.

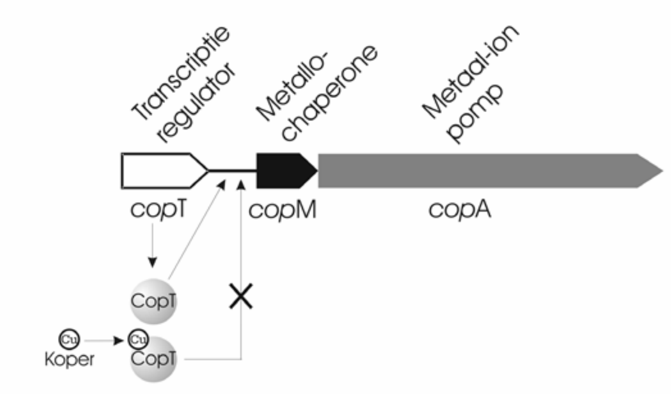
Ten strijde tegen zware metalen

Zoals eerder al werd aangehaald zijn sommige archaea in staat om extreme metaalconcentraties te weerstaan. In ***hoofdstuk 4*** en ***5*** wordt een nieuw mechanisme beschreven dat waarschijnlijk betrokken is bij de resistentie tegen hoge concentraties aan koper en mogelijk andere zware metalen in archaeale soorten. In ***hoofdstuk 4*** word beschreven hoe door archaeale genoomsequenties te vergelijken een operon werd ontdekt bestaande uit genen coderend voor eiwitten die betrokken zijn bij metaal resistentie. Een van deze genen codeert voor een moleculaire pomp (CopA) die waarschijnlijk betrokken is bij het uit de cel pompen van metaal-ionen. Tevens werden er genen ontdekt die coderen voor eiwitten die waarschijnlijk betrokken zijn bij het wegvangen van metaalionen (CopM, een metallochaperone) en bij de regulatie van de transcriptie van deze genen (CopT). Na nauwkeurige inspectie van deze eiwitten bleken deze een nieuw type metaalbindend domein te bevatten, het TRASH domein. Uit een uitgebreide analyse van andere genoomsequenties bleek dat dit domein ook voor te komen in andere eiwitten die betrokken zijn bij metaalresistentie.

METALLOCHAPERONE — Een klein eiwit dat in staat is om specifieke metaal ionen te binden en deze vervolgens door te geven aan een ander eiwit of enzym, bijvoorbeeld een metaal ionen pomp.

PROMOTOR — DNA element dat zich meestal stroomopwaarts van een gen bevindt en dat betrokken is bij de regulatie (het aan- dan wel uitzetten) van de expressie van dit gen. De regulatie wordt meestal gestuurd door specifieke transcriptie regulatoren, die op dit element kunnen binden. Deze regulatoren kunnen een positief (activering) of een negatief effect (repressie) hebben op de expressie van een gen.

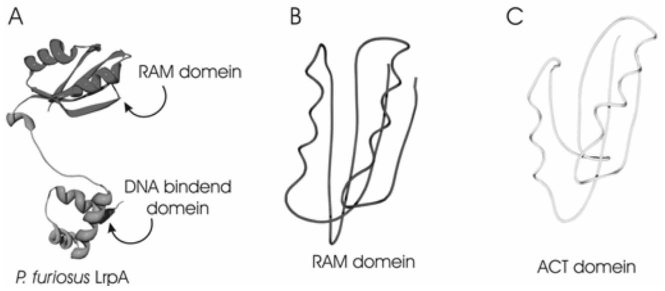
In ***hoofdstuk 5*** staan de resultaten beschreven van experimenteel werk dat verricht is aan het hiervoor genoemde transcriptie regulator CopT uit het archeon *Sulfolobus solfataricus*. Het CopT eiwit bevat naast een DNA-bindend domein ook een al eerder genoemd TRASH domein, dat waarschijnlijk betrokken is bij het meten van aan- of afwezigheid van bepaalde metaalionen in de cel. In *S. solfataricus* liggen de genen die coderen voor het metallochaperone (CopM) en de metaalionen pomp (CopA) geclusterd in een operon, *cop*MA (Figuur 10.4). Uit DNA-bindingsexperimenten blijkt dat CopT op enkele plaatsen bindt op de *cop*MA promotor. In de aanwezigheid van koper ionen blijft deze binding minder efficiënt te zijn. Het lijkt er dus op dat CopT een moleculaire switch is die transcriptie van de *cop*MA genen onderdrukt in afwezigheid van koper. Deze repressie wordt opgeheven als er koper ionen aanwezig zijn (Figuur 10.4). In dit hoofdstuk wordt voor het eerst gerapporteerd over een koper-geïnduceerd resistentie mechanisme in archaea en de resultaten die hier beschreven worden, impliceren een vooraanstaande rol voor het TRASH domein in metaal resistentie in archaea.



Figuur 10.4 Het regulatiemechanisme van het *cop* cluster in *S. solfataricus*. Het *cop*T gen codeert voor een transcriptie regulator (CopT), die bindt aan de promotor van het *cop*MA operon. Deze binding voorkomt dat de *cop*MA genen worden overgeschreven (repressie van transcriptie). Als er koper ionen binden aan CopT, dan kan CopT niet meer aan de promotor van het *cop*MA operon binden en worden de genen overgeschreven (derepressie).

Reguleren is regeren…

In ***hoofdstuk 6*** is dieper ingegaan op het vergelijken van eiwitstructuren in plaats van genomen. Het bestuderen van de structuur of vouwing van eiwitten kan soms waardevolle informatie opleveren over de functie van een eiwit. Het startpunt van dit onderzoek was een transcriptie rgulator van het hyperthermofiele archaeaon *P. furiosus*, LrpA. LrpA behoort tot de Lrp familie van transcriptie regulatoren waarvan bekend dat ze betrokken zijn bij regulatie van aminozuur metabolisme. Lrp eiwitten komen in zowel bacteriële als archaeale soorten voor. Recentelijk is van LrpA de 3D-structuur opgehelderd, waaruit bleek dat het bestaat uit een DNA-bindend domein en een domein waarvan de functie niet bekend was. Na zorgvuldige structurele analyse bleek dat de 3D vouwing van dit onbekende domein erg veel lijkt op de 3D-structuur van ander domein, het ACT domein (Figuur 10.5).



Figuur 10.5 (a) 3D structuur van het LrpA van *P. furiosus*, waarbij duidelijk 2 aparte domeinen te herkennen zijn: een DNA-bindend domein (onder) en het RAM domein. De 3-dimensionale vouwing van het RAM domein **(b)** lijkt erg veel op die van het functioneel verwante ACT domein **(c)**.

Het ACT domein is betrokken bij de regulatie van het aminozuur metabolisme. De overeenkomst in vouwing (‘structurele homologie’) tussen het ACT domein en het onbekende domein versterkt het idee dat het laatstgenoemde domein ook betrokken is bij de regulatie van het aminozuur metabolisme. Daarom is het nieuwe domein ‘RAM’ genoemd (naar *R*egulator van *A*minozuur *M*etabolisme). Zowel het ACT als het RAM domein lijkt een regulerend domein te zijn, wat na het binden van een signaalmolecuul (bijvoorbeeld een aminozuur), een respons doorgeven. Deze respons kan weer een bepaald effect hebben op een eiwit, bijvoorbeeld een verandering in DNA-bindingsaffiniteit, of een toe of afname in enzym activiteit. Na analyse van zgn. responsmutaties van functioneel gekarakteriseerde RAM en ACT domeinen kunnen we concluderen dat de plaats waar ACT en RAM hun signaalmolecuul binden verschillend zijn. Het lijkt er dus op dat ACT en RAM evolutionair gezien verwant zijn, maat dat de bindingsplaatsen van hun signaalmolecuul apart van elkaar geëvolueerd zijn. Beide domein lijken echter een analoge rol te spelen in het controleren van essentiële stappen in het aminozuur metabolisme op het niveau van gen expressie en ook van enzym activiteit.

RESPONSMUTATIE — Een mutatie die tot gevolg heeft dat bij het betreffende eiwit of enzym na binding van een signaalmolecuul niet meer de gangbare respons optreedt. Bijvoorbeeld een DNA-bindend eiwit dat na binding van zijn signaal molecuul niet meer loslaat van het DNA.

Gaten opvullen

Zoals al eerder is aangehaald, verschilt het metabolisme van archaea nogal met dat van bacteriën en eukaryoten. Reconstructie van metabole routes aan de hand van archaeale genoomsequenties resulteert vaak in incomplete routes. Zo blijkt bijvoorbeeld in archaele genomen geen gen voor te komen dat codeert voor het gangbare enzym dat zorgt voor de omzetting van oxaloactetaat naar phospho*enol*pyruvaat. Deze omzetting vormt een belangrijke schakel tussen de glycolyse enerzijds en de citroenzuurcyclus

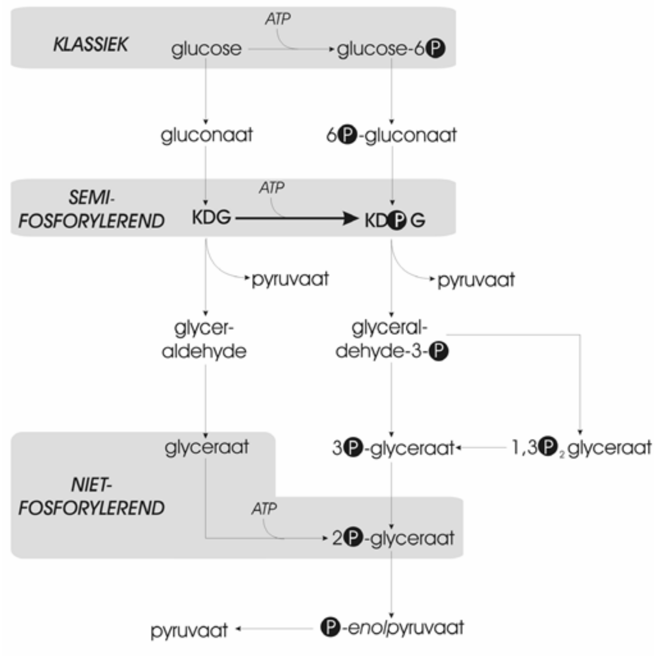
GLYCOLYSE — Centrale afbraakroute in het koolstofmetabolisme waarbij glucose afgebroken wordt tot pyruvaat. Er zijn 2 klassieke typen glycolyse bekend, nl. de Embden-Meyerhoff route en de Entner-Doudoroff route. In de archaeale versies van deze routes komen veel modificaties voor.

anderzijds. In ***hoofdstuk 7*** is met behulp van een bioinformatica aanpak een kandidaat-gen geïdentificeerd dat mogelijk codeert voor deze missende stap in het archaeale metabolisme: het archaeale phospho*enol*pyruvaat carboxylase (atPEPC). In het laboratorium is vervolgens bevestigd dat het geïdentificeerde gen inderdaad codeert voor de missende stap in archaea. Uit een vergelijking van het atPEPC en het klassieke PEPC blijkt dat het archaele enzym een stuk kleiner is en dat het een aantal domeinen mist die te maken hebben bij de regulatie van de enzymactiviteit. Uit de laboratoriumexperimenten blijkt het atPEPC inderdaad een aantal allostere eigenschappen te missen in vergelijking met het klassieke PEPC. De identificatie van de archaele PEPC versie met zijn karakteristieke eigenschappen is het zoveelste voorbeeld van de veelzijdigheid van de enzymen van het centrale koolstofmetabolisme in archaea.

Een variatie op een variatie op een thema

Ook in het laatste hoofdstuk van dit proefschrift, ***hoofdstuk 8***, staat het centrale koolstofmetabolisme van archaea weer centraal. De afbraak van glucose, een centrale speler in het koolstofmetabolisme, wordt glycolyse genoemd. Er zijn 2 verschillende type glycolyse bekend, namelijk de Embden-Meyerhoff (EM) pathway en de Entner-Doudoroff (ED) route. Uit eerder onderzoek aan het metabolisme van archaea is gebleken dat in hun glycolyse een groot aantal variaties voorkomen op de standaard versies van deze metabole routes. Een aantal archaea, waaronder *S. solfataricus* en *Thermoproteus tenax*, gebruiken een afwijkende vorm van de ED route. Op basis van onderzoek dat in de jaren 80 verricht is, werd een afwijkende ED route voorgesteld in deze organismen waarbij de metabolieten niet voorzien worden van een fosfaatgroep ('gefosforyleerd') zoals bij de gangbare ED route: de niet-fosforylerdende ED (nED) route (Figuur 10.6). Het was dan ook enigszins verrassend dat tijdens een analyse van glycoyse genen in archaeale genomen een gen werd aangetroffen dat codeert voor een 'suikerkinase', dat geclusterd lag op het genoom samen met andere genen die coderen voor nED route enzymen. Op basis van dit gegeven zijn een aantal laboratorium experimenten ontworpen om uit te zoeken of er toch

FOSFORYLEREN — Het koppelen van een energierijke fosfaatgroep aan een molecuul (metaboliet of eiwit). Deze koppeling wordt meestal enzymatisch tot stand gebracht door een kinase, waarbij een energierijke fosfaatgroep van een ATP molecuul wordt gekoppeld aan het betreffende molecuul.



Figuur 10.6 Overzicht van de verschillende modificaties die bestaan van de Entner-Doudoroff route. Naast de klassieke route (*boven*), kunnen we de semi-fosforylerende (*midden*) en de niet-fosforylerende ED route (*onder*) onderscheiden, die ieder gekenmerkt worden door het punt in de route waar substraat fosforylereing plaatsvindt (*grijze* kaders). De kenmerkende fosforyleringsstap voor de semi-fosforylerende ED route door het KDG kinase is *dikgedrukt* weergegeven. Fosfaatgroepen zijn met *zwarte* cirkels aangegeven.

fosforylering plaatsvindt in de voorgestelde nED route. Uit dit onderzoek is onder andere gebleken dat het 'suikerkinase' inderdaad een metaboliet (KDG) van de nED route kan voorzien van een fosfaatgroep, resulterend in KDPG. Verder bleek dat een ander essentieel enzym van de nED route, het KDG aldolase, ook in staat om KDPG te splitsen. De resultaten die in dit hoofdstuk worden gepresenteerd duiden er op dat er naast de voorgestelde nED route in deze archaea dus ook een semi-fosforylerende ED route actief is (Figuur 10.6).

Tot slot...

In deze samenvatting heb ik jullie ervan proberen te overtuigen dat archaea fascinerende organismen zijn, met soms verrassende eigenschappen. We beginnen nu pas inzicht te krijgen in de onderliggende biologie van deze eigenschappen, doch keer op keer blijven archaea ons verrassen met hun markante veelzijdigheid. In dit proefschrift is getracht om door archaeale genoomsequenties te analyseren enkele interessante aspecten van hun intrigerende biologie te ontrafelen.

Appendix I

Colour figures

Appendix I

Figure 3.1

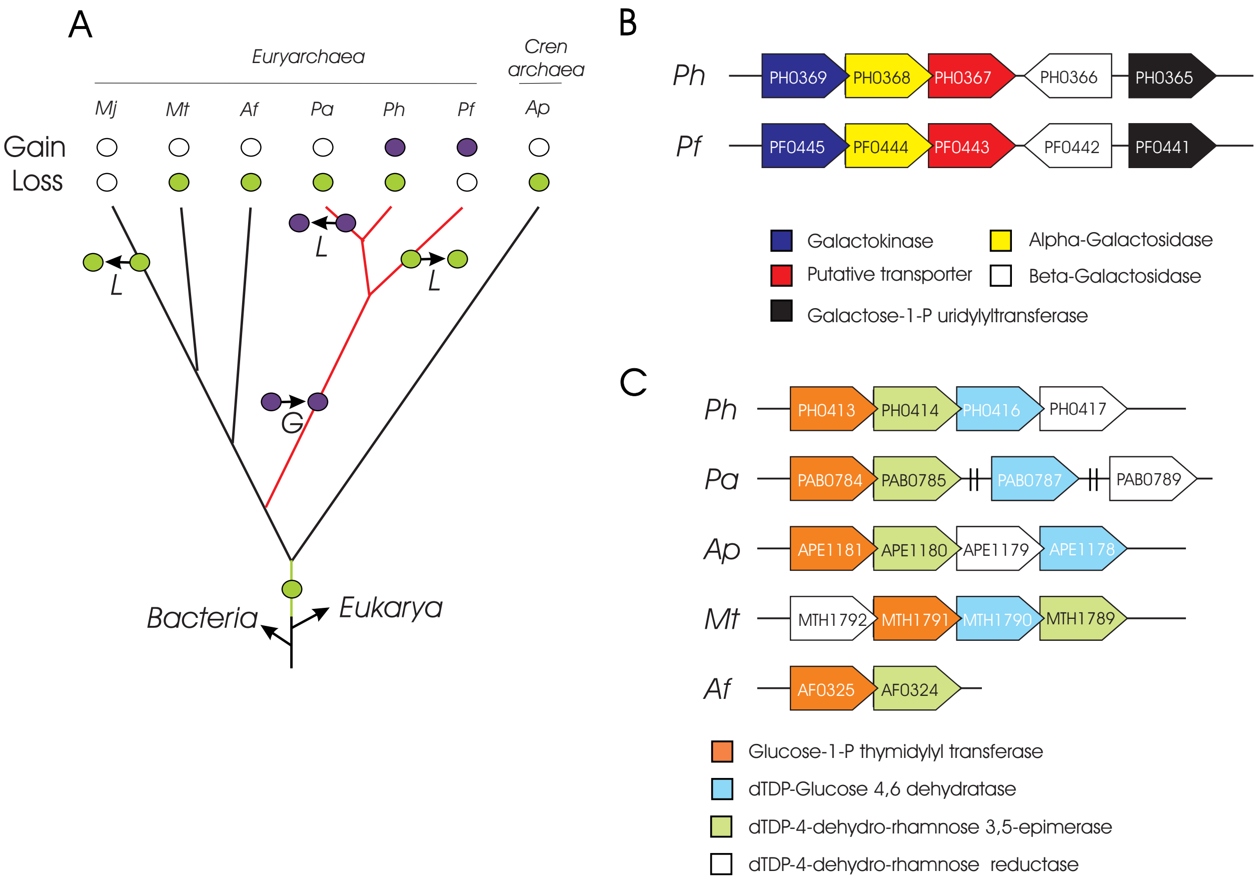


Figure 3.1 Gain and loss of genes in *Pyrococcus*. **(a)** The prediction of gains and losses of genes was based on their phylogenetic distribution, using a genome-based phylogenetic tree[5]. A pyrococcal gain [purple circles, corresponding with cluster in (b)] was defined as a gene that is present in at least one of the *Pyrococci* but has no detectable orthologous counterparts in the other Archaea, thereby selecting genes that were potentially horizontally transferred, duplicated or invented in the pyrococcal branch (*red*). A pyrococcal loss was defined as a gene that has an ortholog in *Aeropyrum pernix* and in at least one of the sequenced Euryarchaea (Af, Mt, Mj), but has no ortholog in one or more *Pyrococci*. *Green* circles correspond to the evolutionary scenario for the example described in (c). Using these conditions, genes were selected that were probably present in the last common archaeal ancestor (*green* branch) and were lost in the pyrococcal branch (*red*). Whether genes were shared between archaeal taxa was determined using an operational definition of orthology. Genes were orthologous when Smith–Waterman-based sequence comparisons[6] between pairs of genes of two genomes displayed highest, significant (e <0.01) bi-directional best hits including the possibility of fission and fusion[1]. **(b)** The presence of a galactoside utilization cluster in both *P. furiosus* and *P. horikoshii* and the absence of these genes in the other Archaea is an example of a gain of functionally interacting genes in these *Pyrococci*. Notice that the *Pyrococci* themselves are phylogenetically too close to consider conservation of gene order among them a significant indication for functional interaction; the galactokinase and the galactose-1-phosphate uridylyl transferase are clustered in numerous Bacteria. **(c)** The presence of a polymeric sugar biosynthesis gene cluster that is probably involved in cell wall biosynthesis in *A. pernix*, *P. abyssi*, *P. horikoshii* and the other Euryarchaeota, but not in *P. furiosus* is an example of a loss event of functionally interacting genes in *P. furiosus*. Notice that the gene cluster is partly lost in *A. fulgidus*. Clusters of functionally interacting genes were detected using STRING (http://www.bork.EMBL-Heidelberg.DE/STRING/). Abbreviations: L, loss event; G, gain event; Mt, *Methanobacterium thermoautotrophicum*; Mj, *Methanococcus jannaschii*; Af, *Archaeoglobus fulgidus*; Ph, *Pyrococcus horikoshii*; Pa, *Pyrococcus abyssi*; Pf, *Pyrococcus furiosus*; Ap, *Aeropyrum pernix*. For availability of genomic coding sequences see http://www.TIGR.org, except for *P. furiosus*[7] (see http://comb5156.umbi.umd.edu/genemate/). To prevent that genes that are annotation artefacts are counted as gains here we only count pyrococcal genes for which we could detect at least one homolog or that are at least 150 amino acids long. Furthermore, all genes that were detected in only a subset of the *Pyrococci* were searched for in the DNA of the other *Pyrococci* using TBLASTN[8], and if found, added to the predicted genes in that genome. For an additional comparison of the pyrococcal genomes, see[9].

Figure 4.1

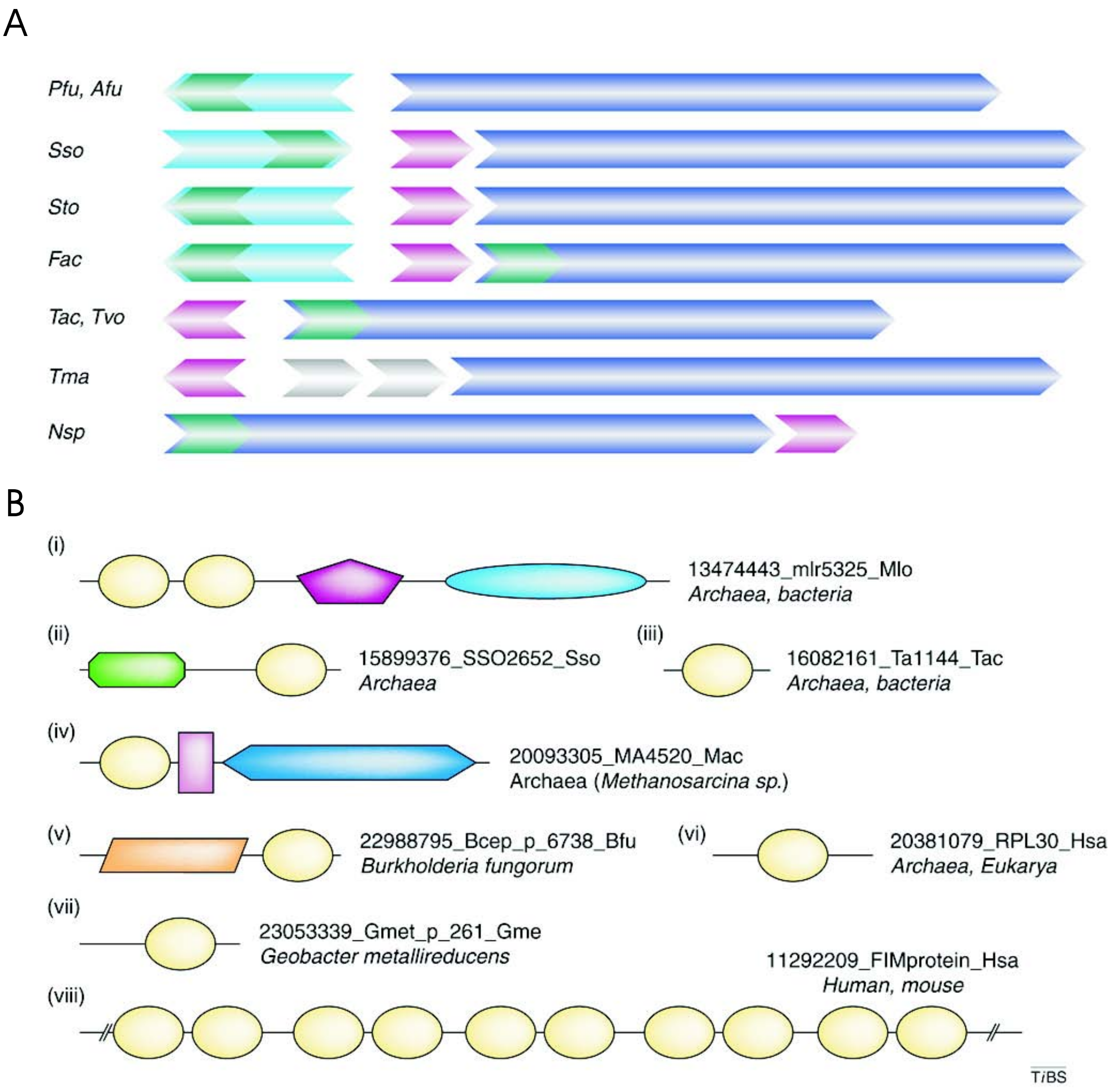
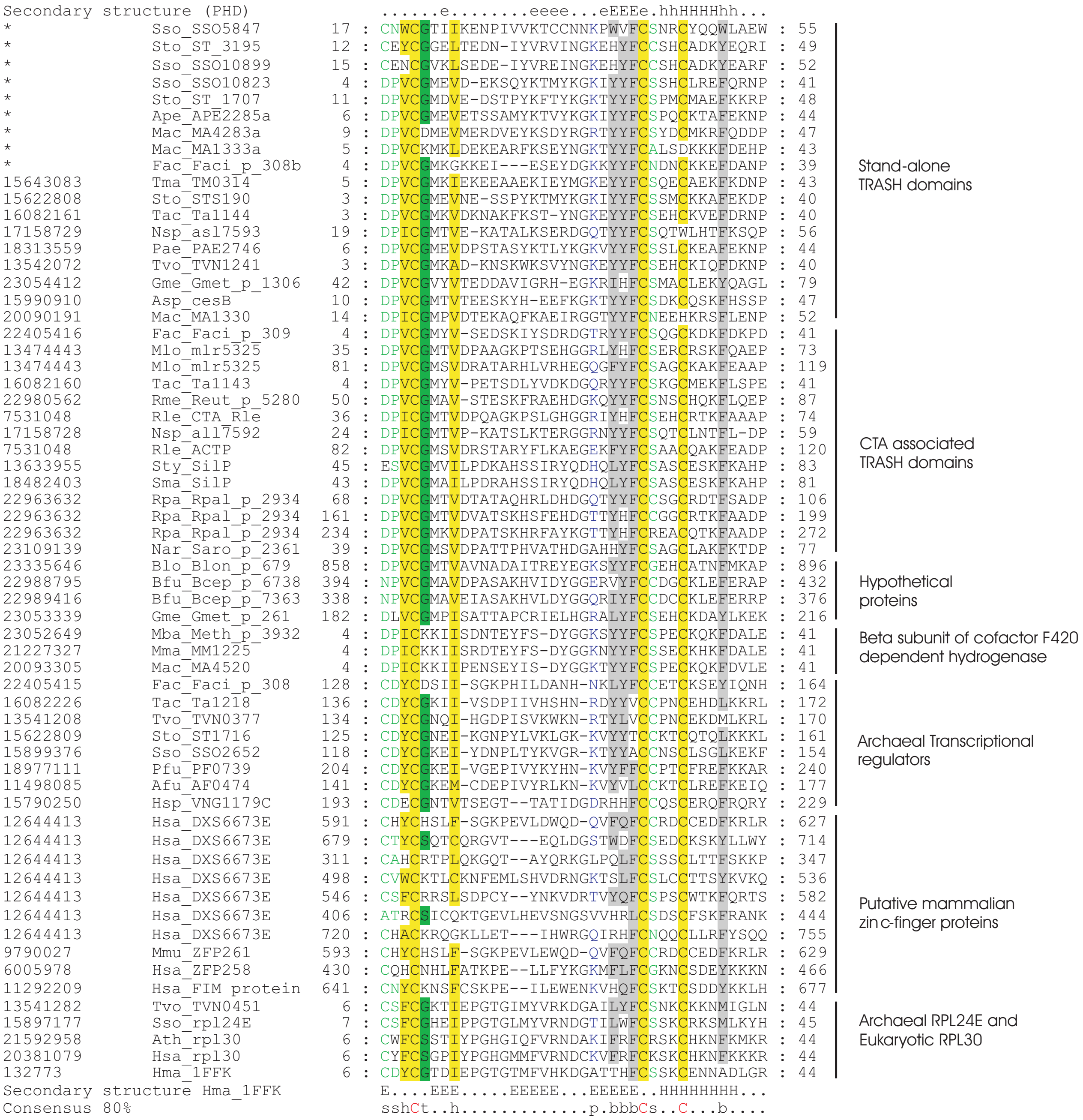

A
Pfu, Afu
Sso
Sto
Fac
Tac, Tvo
Tma
Nsp
B
(i)
13474443_mlr5325_Mlo
Archaea, bacteria
(ii)
15899376_SSO2652_Sso
Archaea
(iii)
16082161_Ta1144_Tac
Archaea, bacteria
(iv)
20093305_MA4520_Mac
Archaea (Methanosarcina sp.)
(v)
22988795_Bcep_p_6738_Bfu
Burkholderia fungorum
(vi)
20381079_RPL30_Hsa
Archaea, Eukarya
(vii)
23053339_Gmet_p_261_Gme
Geobacter metallireducens
(viii)
11292209_FIMprotein_Hsa
Human, mouse
TiBS

Figure 4.1 Physical associations of the TRASH domain. **(a)** Operon and divergon structure of genes encoding TRASH-containing copper chaperones, copper-transporting ATPases (*dark blue*) and transcriptional regulators (*light blue*) are shown. Genes encoding stand-alone TRASH domains (copper chaperones) are indicated in *pink*, TRASH domains of copper-transporting ATPases and transcriptional regulators are indicated in *green*. Unrelated genes are *gray*, arrows are not drawn to scale. The following genes are shown: Pfu: PF0739, PF0740; Afu: AF0474, AF0473; Sso: SSO2652, SSO10823*, SSO2651; Sto: ST1716, STS190, ST1715; Fac: Faci_p_308, Faci_p_308b*, Faci_p_309; Tac: Ta1144, Ta1143; Tvo: TVN6226, TVN6221; Tma: TM0314, TM0317; Nsp: asl7592, asl7593. Genes indicated with an asterisk were ignored during genome annotation and are available at http://www.ftns.wau.nl/micr/bacgen/TRASH/stand-alone.htm. **(b)** Domain architectures of TRASH-containing proteins. Proteins are approximately drawn to scale and denoted by their GenBank Identifier, followed by the species abbreviation and gene name. In addition, the phyletic distribution of the proteins is displayed in italics. Domain architectures were analyzed using SMART[12] and Pfam[13]. The TRASH-containing proteins are subdivided into the following classes: (*i*) P-type cation-transporting ATPases, typically containing one to three N-terminally fused TRASH domains (*yellow*), an E1-E2_ATPase domain (PF00122; *pink*) and a hydrolase domain (PF00702; *light blue*); (*ii*) transcriptional regulators containing an N-terminal helix–turn–helix motif (HTH_ASNC; *light green*) and a C-terminal TRASH domain; (*iii*) metallochaperone-like proteins, consisting of a single TRASH domain only; (*iv*) β-subunit of coenzyme F420-dependent hydrogenase, containing an N-terminal TRASH domain, followed by a ferredoxin domain (Fer4; PF00037; *purple*) and the C-terminal F420 dependent hydrogenase catalytic domain (*dark blue*); (v) uncharacterized protein, containing an N-terminal domain (PF02625) that is related to xanthine and CO dehydrogenase maturation factor (COG1975) and a C-terminal TRASH domain; (*vi*) Archaeal (RPL24E) and eukaryote (RPL30) ribosomal proteins containing a single N-terminal TRASH domain. (*vii*) hypothetical protein containing a C-terminal TRASH domain; (*viii*) putative mammalian zinc-finger proteins, containing multiple tandem repeated TRASH domains. Abbreviations: Afu, *Archaeoglobus fulgidus*; Asp, *Acinetobacter* sp. BW3; Ath, *Arabodopsis thaliana*; Avi, *Azotobacter vinelandii*; Bfu, *Burkholderia fungorum*; Blo, *Bifidobacterium longum*; Fac, *Ferroplasma acidarmanus*; Gme, *Geobacter metallireducens*; Hma, *Haloarcula marismortui*; Hsa, *Homo sapiens*; Hsp, *Halobacterium* sp. NRC-1; Mac, *Methanosarcina acetivorans*; Mba, *Methanosarcina barkeri*; Mlo, *Mesorhizobium loti*; Mma, *Methanosarcina mazei*; Mmu, *Mus musculus*; Nar, *Novosphingobium aromaticivorans*; Nsp, *Nostoc* sp. PCC 7120; Pae, *Pyrobaculum aerophilum*; Pfu, *Pyrococcus furiosus*; Rle, *Rhizobium leguminosarum*; Rme, *Ralstonia metallidurans*; Rpa, *Rhodopseudomonas palustris*; Sma, *Serratia marcescens*; Sso, *Sulfolobus solfataricus*; Sto, *Sulfolobus tokodaii*; Sty, *Salmonella typhimurium*; Tac, *Thermoplasma acidophilum*; Tma, *Thermotoga maritima*; Tvo, *Thermoplasma volcanium*.

Figure 4.2

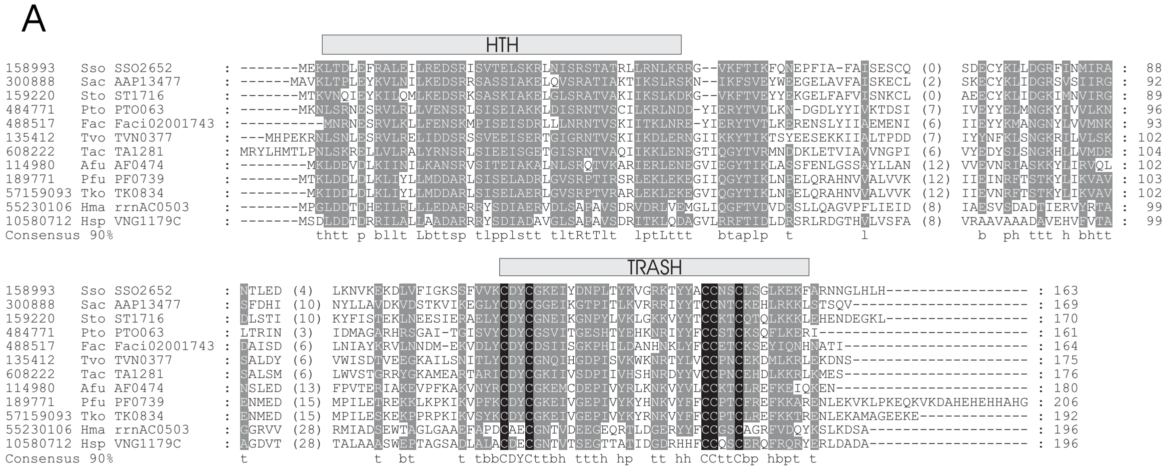


```
Secondary structure (PHD)                  ......e........eeee...eEEEe.hhHHHHhh...
*                Sso_SSO5847        17 : CNWCGTIIKENPIVVKTCCNNKPWVFCSNRCYQQWLAEW : 55
*                Sto_ST_3195        12 : CEYCGGELTEDN-IYVRVINGKEHYFCCSHCADKYEQRI : 49
*                Sso_SSO10899       15 : CENCGVKLSEDE-IYVREINGKEHYFCCSHCADKYEARF : 52
*                Sso_SSO10823        4 : DPVCGMEVD-EKSQYKTMYKGKIYYFCSSHCLREFQRNP : 41
*                Sto_ST_1707        11 : DPVCGMDVE-DSTPYKFTYKGKTYYFCSPMCMAEFKKRP : 48
*                Ape_APE2285a        6 : DPVCGMEVETSSAMYKTVYKGKIYYFCSPQCKTAFEKNP : 44
*                Mac_MA4283a         9 : DPVCDMEVMERDVEYKSDYRGRTYYFCSYDCMKRFQDDP : 47
*                Mac_MA1333a         5 : DPVCKMKLDEKEARFKSEYNGKTYYFCALSDKKKFDEHP : 43
*                Fac_Faci_p_308b     4 : DPVCGMKGKKEI---ESEYDGKKYYFCNDNCKKEFDANP : 39
15643083         Tma_TM0314          5 : DPVCGMKIEKEEAAEKIEYMGKEYYFCSQECAEKFKDNP : 43
15622808         Sto_STS190          3 : DPVCGMEVNE-SSPYKTMYKGKIYYFCSSMCKKAFEKDP : 40
16082161         Tac_Ta1144          3 : DPVCGMKVDKNAKFKST-YNGKEYYFCSEHCKVEFDRNP : 40
17158729         Nsp_asl7593        19 : DPICGMTVE-KATALKSERDGQTYYFCSQTWLHTFKSQP : 56
18313559         Pae_PAE2746         6 : DPVCGMEVDPSTASYKTLYKGKVYYFCSSLCKEAFEKNP : 44
13542072         Tvo_TVN1241         3 : DPVCGMKAD-KNSKWKSVYNGKEYYFCSEHCKIQFDKNP : 40
23054412         Gme_Gmet_p_1306    42 : DPVCGVYVTEDDAVIGRH-EGKRIHFCSMACLEKYQAGL : 79
15990910         Asp_cesB           10 : DPVCGMTVTEESKYH-EEFKGKTYYFCSDKCQSKFHSSP : 47
20090191         Mac_MA1330         14 : DPICGMPVDTEKAQFKAEIRGGTYYFCNEEHKRSFLENP : 52
22405416         Fac_Faci_p_309      4 : DPVCGMYV-SEDSKIYSDRDGTRYYFCSQGCKDKFDKPD : 41
13474443         Mlo_mlr5325        35 : DPVCGMTVDPAAGKPTSEHGGRLYHFCSERCRSKFQAEP : 73
13474443         Mlo_mlr5325        81 : DPVCGMSVDRATARHLVRHEGQGFYFCSAGCKAKFEAAP : 119
16082160         Tac_Ta1143          4 : DPVCGMYV-PETSDLYVDKDGQRYYFCSKGCMEKFLSPE : 41
22980562         Rme_Reut_p_5280    50 : DPVCGMAV-STESKFRAEHDGKQYYFCSNSCHQKFLQEP : 87
7531048          Rle_CTA_Rle        36 : DPICGMTVDPQAGKPSLGHGGRIYHFCSEHCRTKFAAAP : 74
17158728         Nsp_all7592        24 : DPICGMTVP-KATSLKTERGGRNYYFCSQTCLNTFL-DP : 59
7531048          Rle_ACTP           82 : DPVCGMSVDRSTARYFLKAEGEKFYFCSAACQAKFEADP : 120
13633955         Sty_SilP           45 : ESVCGMVILPDKAHSSIRYQDHQLYFCSASCESKFKAHP : 83
18482403         Sma_SilP           43 : DPVCGMAILPDRAHSSIRYQDHQLYFCSASCESKFKAHP : 81
22963632         Rpa_Rpal_p_2934    68 : DPVCGMTVDTATAQHRLDHDGQTYYFCCSGCRDTFSADP : 106
22963632         Rpa_Rpal_p_2934   161 : DPVCGMTVDVATSKHSFEHDGTTYHFCCGGCRTKFAADP : 199
22963632         Rpa_Rpal_p_2934   234 : DPVCGMKVDPATSKHRFAYKGTTYHFCREACQTKFAADP : 272
23109139         Nar_Saro_p_2361    39 : DPVCGMSVDPATTPHVATHDGAHHYFCSAGCLAKFKTDP : 77
23335646         Blo_Blon_p_679    858 : DPVCGMTVAVNADAITREYEGKSYYFCGEHCATNFMKAP : 896
22988795         Bfu_Bcep_p_6738   394 : NPVCGMAVDPASAKHVIDYGGERVYFCCDGCKLEFERAP : 432
22989416         Bfu_Bcep_p_7363   338 : NPVCGMAVEIASAKHVLDYGGQRIYFCCDCCKLEFERRP : 376
23053339         Gme_Gmet_p_261    182 : DLVCGMPISATTAPCRIELHGRALYFCSEHCKDAYLKEK : 216
23052649         Mba_Meth_p_3932     4 : DPICKKIISDNTEYFS-DYGGKSYYFCSPECKQKFDALE : 41
21227327         Mma_MM1225          4 : DPICKKIISRDTEYFS-DYGGKNYYFCSSECKHKFDALE : 41
20093305         Mac_MA4520          4 : DPICKKIIPENSEYIS-DYGGKTYYFCSPECKQKFDVLE : 41
22405415         Fac_Faci_p_308    128 : CDYCDSII-SGKPHILDANH-NKLYFCCETCKSEYIQNH : 164
16082226         Tac_Ta1218        136 : CDYCGKII-VSDPIIVHSHN-RDYYVCCPNCEHDLKKRL : 172
13541208         Tvo_TVN0377       134 : CDYCGNQI-HGDPISVKWKN-RTYLVCCPNCEKDMLKRL : 170
15622809         Sto_ST1716        125 : CDYCGNEI-KGNPYLVKLGK-KVYYTCCKTCQTQLKKKL : 161
15899376         Sso_SSO2652       118 : CDYCGKEI-YDNPLTYKVGR-KTYYACCNSCLSGLKEKF : 154
18977111         Pfu_PF0739        204 : CDYCGKEI-VGEPIVYKYHN-KVYFFCCPTCFREFKKAR : 240
11498085         Afu_AF0474        141 : CDYCGKEM-CDEPIVYRLKN-KVYVLCCKTCLREFKEIQ : 177
15790250         Hsp_VNG1179C      193 : CDECGNTVTSEGT--TATIDGDRHHFCCQSCERQFRQRY : 229
12644413         Hsa_DXS6673E      591 : CHYCHSLF-SGKPEVLDWQD-QVFQFCCRDCCEDFKRLR : 627
12644413         Hsa_DXS6673E      679 : CTYCSQTCQRGVT---EQLDGSTWDFCSEDCKSKYLLWY : 714
12644413         Hsa_DXS6673E      311 : CAHCRTPLQKGQT--AYQRKGLPQLFCSSSCLTTFSKKP : 347
12644413         Hsa_DXS6673E      498 : CVWCKTLCKNFEMLSHVDRNGKTSLFCSLCCTTSYKVKQ : 536
12644413         Hsa_DXS6673E      546 : CSFCRRSLSDPCY--YNKVDRTVYQFCSPSCWTKFQRTS : 582
12644413         Hsa_DXS6673E      406 : ATRCSICQKTGEVLHEVSNGSVVHRLCSDSCFSKFRANK : 444
12644413         Hsa_DXS6673E      720 : CHACKRQGKLLET---IHWRGQIRHFCNQQCLLRFYSQQ : 755
9790027          Mmu_ZFP261        593 : CHYCHSLF-SGKPEVLEWQD-QVFQFCCRDCCEDFKRLR : 629
6005978          Hsa_ZFP258        430 : CQHCNHLFATKPE--LLFYKGKMFLFCGKNCSDEYKKKN : 466
11292209         Hsa_FIM protein   641 : CNYCKNSFCSKPE--ILEWENKVHQFCSKTCSDDYKKLH : 677
13541282         Tvo_TVN0451         6 : CSFCGKTIEPGTGIMYVRKDGAILYFCSNKCKKNMIGLN : 44
15897177         Sso_rpl24E          7 : CSFCGHEIPPGTGLMYVRNDGTILWFCSSKCRKSMLKYH : 45
21592958         Ath_rpl30           6 : CWFCSSTIYPGHGIQFVRNDAKIFRFCRSKCHKNFKMKR : 44
20381079         Hsa_rpl30           6 : CYFCSGPIYPGHGMMFVRNDCKVFRFCKSKCHKNFKKKR : 44
132773           Hma_1FFK            6 : CDYCGTDIEPGTGTMFVHKDGATTHFCSSKCENNADLGR : 44
Secondary structure Hma_1FFK               E....EEE.....EEEEE...EEEEE..HHHHHHHH...
Consensus 80%                              sshCt..h.............p.bbbCs..C...b....
```

Stand-alone TRASH domains
CTA associated TRASH domains
Hypothetical proteins
Beta subunit of cofactor F420 dependent hydrogenase
Archaeal Transcriptional regulators
Putative mammalian zinc-finger proteins
Archaeal RPL24E and Eukaryotic RPL30

Figure 4.2 Multiple sequence alignment of TRASH domains that were identified using iterative PSI–BLAST searches and Hidden Markov profile searches[18]. The 80% consensus shown below the alignment was derived using the following amino-acid classes: hydrophobic [*h* (A,L,I,C,V,M,Y,F,W); *yellow* shading]; small [*s* (A,C,D,G,N,P,S,T,V); *green*] and the tiny subset of the small class [*t* (G,A,S); *green* shading]; polar [*p* (C,D,E,H,K,N,Q,R,S,T); *blue* text] and big [*b* (E,F,I,L,M,Q,R,W,Y); *gray* shading]. Completely conserved cysteine residues within the hydrophobic class are shown in red in the consensus. The secondary structure of TRASH was predicted using PHD[15] and is depicted above the alignment; it is in agreement with the known structure of the *Haloarcula marismortui* RPL24E[16], which is depicted under the alignment. β-strands and α-helices are represented by *e* and *h*, respectively; uppercase indicates the prediction has an accuracy >82%, lowercase represents an accuracy >72%. The limits of the domains are indicated by the position numbers on each side and the sequences are denoted by their GenBank Identifier, followed by the species abbreviation and gene name. Stand-alone versions of the TRASH domain that were missed in the initial genome annotation are indicated with an asterisk (protein sequences are available at http://www.ftns.wau.nl/micr/bacgen/TRASH/stand-alone.htm). This multiple sequence alignment (alignment number ALIGN_000512) has been deposited with the European Bioinformatics Institute (ftp://ftp.ebi.ac.uk/pub/databases/embl/align/ALIGN_000512).

Figure 5.2



```
158993   Sso SSO2652      : -------MEKLTDLEFRALEILREDSRISVTELSKRLNISRSTATRLLRNLKRRG--VKFTIKFQNEPFIA-FAISESCQ (0)  SDECYKLLDGRFINMIRA :  88
300888   Sac AAP13477     : ------MAVKLTPLEYKVLNILKEDSRRSASSIAKELQVSRATIAKTIKSLRSKN--IKFSVEYWEEGELAVFAISKECL (2)  SKECYKLIDGRSVSIIRG :  92
159220   Sto ST1716       : -------MTKVNQIEYKILQMLKEDSRKSASKIAKELGLSRATVAKIIKSLKDKG--VKFTVEYYEKGELFAFVISNKCL (0)  AEECYKLIDGKIMNVIRG :  89
484771   Pto PTO063       : -------MKNLSRNESRVLRLLVENSRLSISEIAKKLDISRNTVSCILKRLNDD-YIERYTVDLKN-DGDLYYIVKTDSI (7)  IVEYYELMNGKYIVVLKN :  96
488517   Fac Faci02001743 : ----------MNRNESRVLKLLFENSKMPISEISDRLLLNRNTVSKIITKLNRE-YIERYTVTLKERENSLYIIAEMENI (6)  IIEYYKMANGNYLVVMNK :  93
135412   Tvo TVN0377      : ---MHPEKRNLSNLESRVLRELIDDSRSSVEEISERTGIGRNTVSKIIKDLERNGIIKKYTIKTSYEESEKKIIALTPDD (7)  IYYNFKLSNGKRILVLSK : 102
608222   Tac TA1281       : MRYLHMTLPNLSKRELLVLRALYRNSRLSIEEISGETGISRNTVAQIIKKLENEGTIQGYTVRMNDDKLETVIAVVNGPI (6)  VYEDYSLSNGKHLLVMDR : 104
114980   Afu AF0474       : --------MKLDEVDLKIINILKANSRVSITEIAKALNLSRQTVKARIERLENEGVIEGYTIKLASSFENLGSSAYLLAN (12) VVEVNRIASKKYLIRVQL : 102
189771   Pfu PF0739       : -------MTKLDDLDLKLIYLLMDDARLSISELADRLGVSRPTIRSRLEKLEKEGIIQGYTIKLNPELQRAHNVVALVVK (12) IIEINRFTSTKYLIKVAV : 103
57159093 Tko TK0834       : --------MKIDDLDLKLIYLLMDDARLSISELAERLGVSRPTVRARLEKLEKEGVIQGYTIKLNPELQRAHNVVALVVK (12) IIEVNRFTSTKYLIKVAV : 102
55230106 Hma rrnAC0503    : -------MPGLDDTDHEILRLLLEDARRRYSDIAERVDLSAPAVSDRVDRLVEMGLIQGFTVDVDRSLLQAGVPFLIEID (8)  IAESVSDADTIERVYRTA :  99
10580712 Hsp VNG1179C     : -------MSDLDDTDRRILALLAADARRRYSDIADAVGLSAPAVSDRITKLQDAGVLRRFTIDLDRSRLRDGTHVLVSFA (8)  VRAAVAAADAVEHVFVTA :  99
Consensus 90%                        thtt p bllt Lbttsp tlpplstt tltRtTlt  lptLttt   btaplp  t        l             b  ph ttt h bhtt

158993   Sso SSO2652      : NTLED (4)  LKNVKEKDLVFIGKSSFVVKCDYCGKEIYDNPLTYKVGRKTYYACCNSCLSGLKEKFARNNGLHLH----------------- : 163
300888   Sac AAP13477     : SFDHI (10) NYLLAVDKVDSTKVIKEGLYCDYCGGEIKGTPITLKVRRRIYYTCCNTCKEHLRKKLSTSQV--------------------- : 169
159220   Sto ST1716       : DLSTI (10) KYFISTEKLNEESIERAELYCDYCGNEIKGNPYLVKLGKKVYYTCCKTCQTQLKKKLEHENDEGKL----------------- : 170
484771   Pto PTO063       : LTRIN (3)  IDMAGARHRSGAI-TGISVYCDYCGSVITGESHTYEHKNRIYYFCCSTCKSQFLKERI------------------------- : 161
488517   Fac Faci02001743 : DAISD (6)  LNIAYKRVLNNDM-EKVDLYCDYCDSIISGKPHILDANHNKLYFCCETCKSEYIQNHNATI---------------------- : 164
135412   Tvo TVN0377      : SALDY (6)  VWISDTVEEGKAILSNITLYCDYCGNQIHGDPISVKWKNRTYLVCCPNCEKDMLKRLEKDNS--------------------- : 175
608222   Tac TA1281       : SALSM (6)  LWVSTGRRYGKAMEARTARICDYCGKIIVSDPIIVHSHNRDYYVCCPNCEHDLKKRLKMES---------------------- : 176
114980   Afu AF0474       : NSLED (13) FPVTERIAKEVPFKAKVNYRCDYCGKEMCDEPIVYRLKNKVYVLCCKTCLREFKEIQKEN----------------------- : 180
189771   Pfu PF0739       : ENMED (15) MPILETREKKLKPKIKVPFKCDYCGKEIVGEPIVYKYHNKVYFFCCPTCFREFKKARENLEKVKLPKEQKVKDAHEHEHHAHG : 206
57159093 Tko TK0834       : ENMED (15) MPILESKEKPPRPKIKVSYKCDYCGKEIVGEPIVYKYRNKVYFFCCPTCLREFKKTRENLEKAMAGEEKE------------- : 192
55230106 Hma rrnAC0503    : GGRVV (28) RMIADSEWTAGLGAAEFAPDCAECGNTVDEEGEQRTLDGERYYFCCGSCAGRFVDQYKSLKDSA------------------- : 196
10580712 Hsp VNG1179C     : AGDVT (28) TALAAASWEPTAGSADLALACDECGNTVTSEGTTATIDGDRHHFCCQSCERQFRQRYERLDADA------------------- : 196
Consensus 90%               t               t  bt     t tbbCDYCttbh ttth hp  tt hh CCttCbp hbpt t
```

B

TRASH domain

```
Secondary structure (PHD pred)     ......e........eeee...eEEEe.hhHHHHhh...
15899376     Sso SSO2652      118 : CDYCGKEI-YDNPLTYKVGR-KTYYACCNSCLSGLKEKF : 154
132773     Hma 1FFK             6 : CDYCGTDIEPGTGTMFVHKDGATTHFCSSKCENNADLGR : 44
Secondary structure (Hma 1FFK)     E....EEE.....EEEEE...EEEEE..HHHHHHHH...
```

HTH domain

```
Secondary structure (PHD pred)     ..HHHHHHHHHHH....HHHHHHHHHH....HHHHHHHHHHHH
158993   Sso SSO2652            4 : LTDLEFRALEILREDSRISVTELSKRLNISRSTATRLLRNLKR : 47
4416541  Pfu 1I1G               2 : IDERDKIILEILEKDARTPFTEIAKKLGISETAVRKRVKALEE : 45
Secondary structure (Pfu 1I1G)     ...HHHHHHHHHHHTTTTT.HHHHHHHHT...HHHHHHHHH..
```

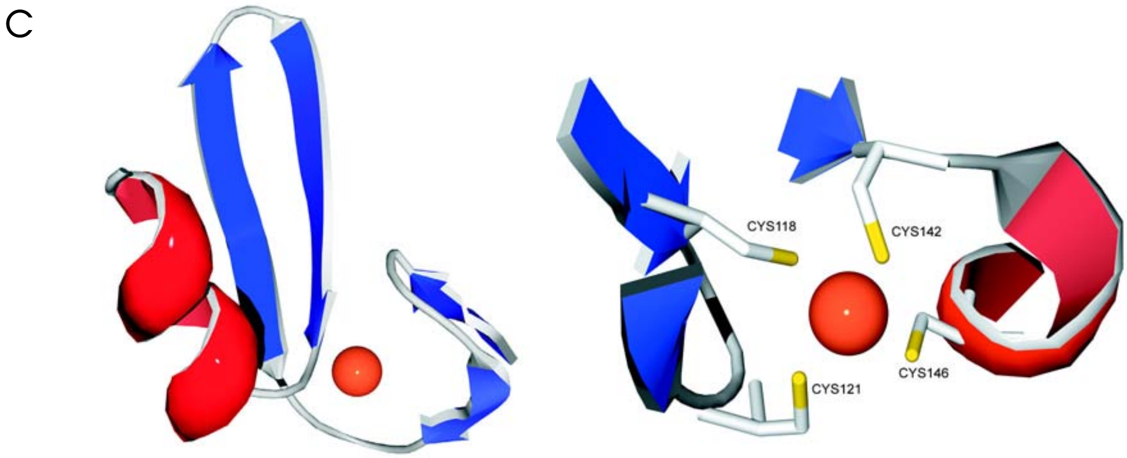
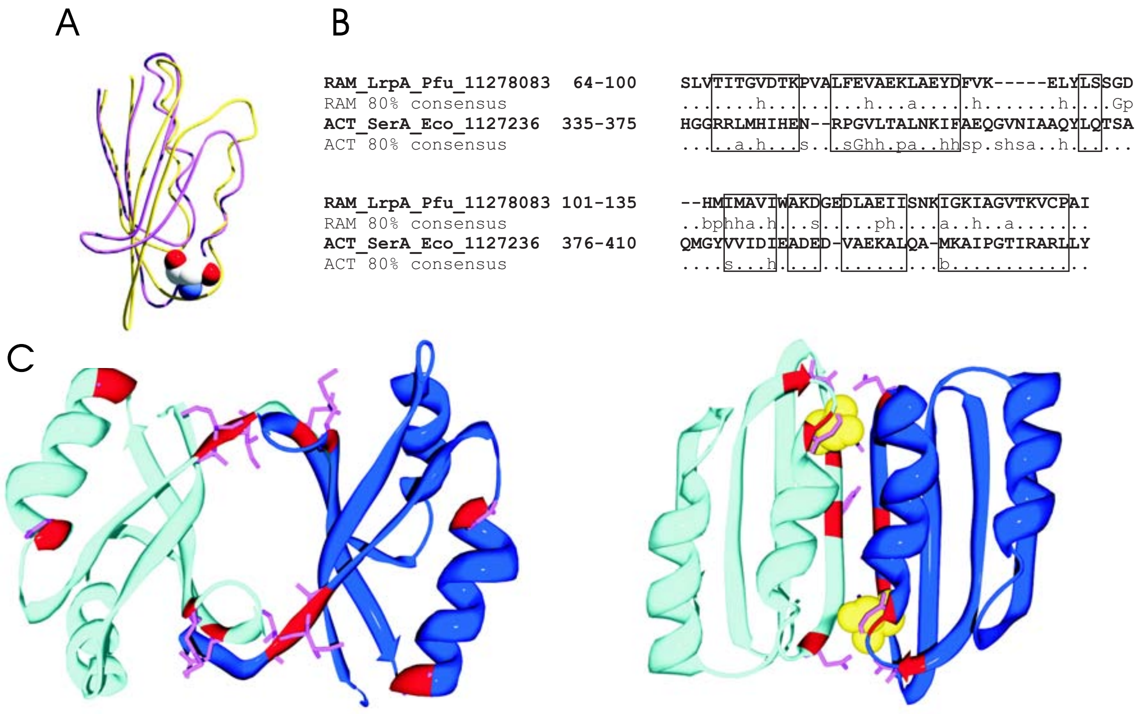


Figure 5.2 Sequence analysis of CopT. **(a)** Multiple alignment of the core of archaeal CopT proteins. The sequences are denoted as in figure 5.1 (except that for *S. acidocaldarius* CopT a locus entry is provided). The positions of the first and the last residue of the aligned region in the corresponding protein are indicated for each sequence. The 95% consensus sequence shown below the alignments was obtained using the following amino acid classes: aromatic (*a*: FYW), small (*s*: GASC), charged (*p*: STEDKRNQH), turn-forming (*t*: ASTDNVGPENRK), big (*b*: FILMVWYKREQ) and hydrophobic (*h*: ACFILMVWYH), of which the aliphatic subset (*l*: ILVA). Residues conserved > 95% are shown in *white* and boxed *gray*. Conserved cysteine residues that have been predicted to be involved in metal binding are shown in white and boxed black. The location of a putative helix-turn-helix motif and the TRASH domain are indicated above the alignment. The numbers within the alignment represent poorly conserved inserts that are not shown. For the CopT sequences of *P. furiosus* (PF0739) and *Halobacterium* sp. (VNG1179C), the transcription starts have been adjusted. **(b)** Structural alignment of the putative HTH domain of *S. solfataricus* CopT and the HTH domain of *P. furiosus* LrpA (*upper* panel), and of the TRASH domain of *S. solfataricus* CopT and the TRASH domain of *H. marismortui* RPL24E (*lower* panel). The sequences are denoted by Gene Identifier (gi) numbers from the GenBank database, species abbreviation and systematic gene numbers, respectively. Known structures of LrpA and RPL24E are depicted under the alignments. The predicted secondary structures of both domains of CopT were generated using PHD[39] and are depicted above the respective alignments. **(c)** Model of TRASH domain of *S. solfataricus* CopT based on homology modeling with the 3D structure of the ribosomal protein RPL24E of *H. marismortui*. *Left* panel: Overview of the modeled treble clef fold. *Right* panel: Close-up of the proposed metal binding site of the TRASH domain, based on the metal binding site of Hm-RPL24E. Cysteine residues that are anticipated to be involved in metal binding are numbered according to amino acid residue numbers of *S. solfataricus* CopT. Species abbreviations: Afu, *Archaeoglobus fulgidus*; Fac: *Ferroplasma acidarmanus*; Hma: *Haloarcula marismortui*; Hsp: *Halobacterium* sp. NRC1; Pto: *Picrophilus torridus*; Pfu: *Pyrococcus furiosus*; Sac: *Sulfolobus acidocaldarius*; Sso: *Sulfolobus solfataricus*; Sto: *Sulfolobus tokodaii*; Tko: *Thermococcus kodakaraensis*; Tac: *Thermoplasma acidophilum*; Tvo: *Thermoplasma volcanium*.

Figure 6.1

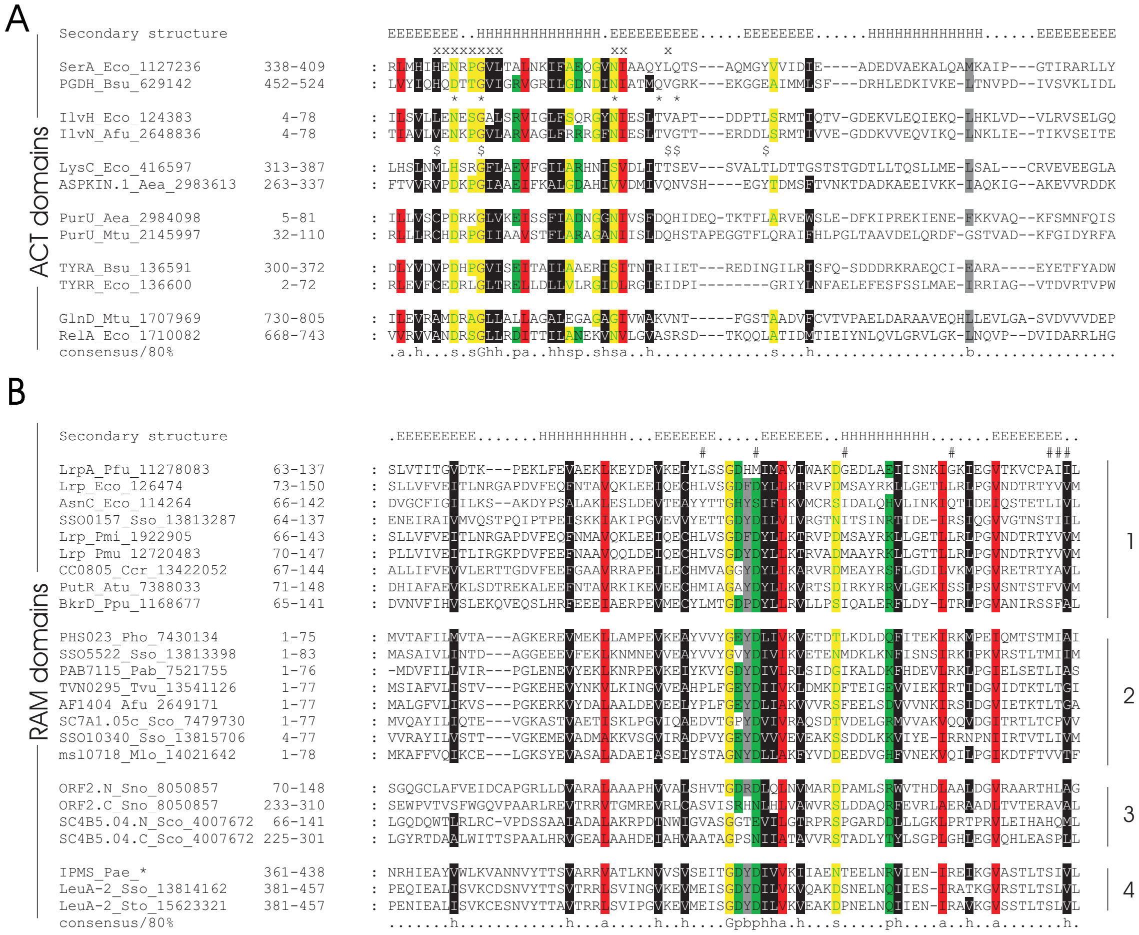


```
RAM_LrpA_Pfu_11278083  64-100     SLVTITGVDTKPVALFEVAEKLAEYDFVK-----ELYLSSGD
RAM 80% consensus                 .......h.......h...a.....h.......h....Gp
ACT_SerA_Eco_1127236  335-375     HGGRRLMHIHEN--RPGVLTALNKIFAEQGVNIAAQYLQTSA
ACT 80% consensus                 .....a.h...s...sGhh.pa..hhsp.shsa..h......

RAM_LrpA_Pfu_11278083 101-135     --HMIMAVIWAKDGEDLAEIISNKIGKIAGVTKVCPAI
RAM 80% consensus                 ..bphha.h...s.....ph....a..h..a.......
ACT_SerA_Eco_1127236  376-410     QMGYVVIDIEADED-VAEKALQA-MKAIPGTIRARLLY
ACT 80% consensus                 ....s...h...............b.............
```

Figure 6.1 (a) Superimposition of the RAM domain of the *P. furiosus* LrpA (*purple*) with the ACT domain of the *E. coli* SerA (*yellow*) showing the strong similarity between the two domains at the structural level (root mean square deviation value 1.8 Å). The superimposition was constructed with 49 -carbon atoms, which composes about 65% of the domains. In addition, the position of the negative effector of SerA ACT domain, serine, was indicated within the superimposed domains. Superimposition and figure were created using the Swiss PDB viewer[10]. **(b)** Structural alignment of the RAM domain of the *P. furiosus* LrpA (residues 64-135) and the ACT domain of the *E. coli* SerA (residues 335-410). In addition, the 80% consensus sequences for RAM and ACT domains are included in the alignment indicating the sequence divergence between the two SMBDs. For abbreviations of the different amino acid classes see figure 2. Matched residues that display a root mean square value of less than 2.5 Å are boxed. The structural alignment was constructed using the Swiss PDB viewer[10] using the "Structural alignment" option. **(c)** Structural comparison of the RAM dimer of *P. furiosus* LrpA (*left*) and ACT dimer of *E. coli* SerA (*right*). The monomers are shown in *cyan* and *blue*, and the ligand response mutations of the RAM domain and the ACT domain corresponding to those that are depicted in figure 2, (a) and (b), respectively, were mapped into the backbones of the respective structures in *red* with *magenta* side chains.

Figure 6.2



```
Secondary structure                EEEEEEEE..HHHHHHHHHHHHHH.EEEEEEEEEE.....EEEEEEEE......HHHHHHHHHHHHH......EEEEEEEEE
                                        xxxxxxxx            xx    x
SerA_Eco_1127236        338-409  : RLMHIHENRPGVLTALNKIFAEQGVNIAAQYLQTS---AQMGYVVIDIE----ADEDVAEKALQAMKAIP---GTIRARLLY
PGDH_Bsu_629142         452-524  : LVYIQHQDTTGVIGRVGRILGDNDINIATMQVGRK---EKGGEAIMMLSF---DRHLEDKIVKE-LTNVPD--IVSVKLIDL
                                          *  *              *    * *
IlvH_Eco_124383            4-78  : ILSVLLENESGALSRVIGLFSQRGYNIESLTVAPT---DDPTLSRMTIQTV-GDEKVLEQIEKQ-LHKLVD--VLRVSELGQ
IlvN_Afu_2648836           4-78  : TIAVLVENKPGVLARVAGLFRRRGFNIESLTVGTT---ERDDLSRMTIVVE-GDDKVVEQVIKQ-LNKLIE--TIKVSEITE
                                        $    $                    $$         $
LysC_Eco_416597         313-387  : LHSLNMLHSRGFLAEVFGILARHNISVDLITTSEV---SVALTLDTTGSTSTGDTLLTQSLLME-LSAL---CRVEVEEGLA
ASPKIN.1_Aea_2983613    263-337  : FTVVRVPDKPGIAAEIFKALGDAHIVVDMIVQNVSH----EGYTDMSFTVNKTDADKAEEIVKK-IAQKIG--AKEVVRDDK

PurU_Aea_2984098           5-81  : ILLVSCPDRKGLVKEISSFIADNGGNIVSFDQHIDEQ-TKTFLARVEWSLE-DFKIPREKIENE-FKKVAQ--KFSMNFQIS
PurU_Mtu_2145997         32-110  : RLLLRCHDRPGIIAAVSTFLARAGANIISLDQHSTAPEGGTFLQRAIFHLPGLTAAVDELQRDF-GSTVAD--KFGIDYRFA

TYRA_Bsu_136591         300-372  : DLYVDVPDHPGVISEITAILAAERISITNIRIIET---REDINGILRISFQ-SDDDRKRAEQCI-EARA----EYETFYADW
TYRR_Eco_136600            2-72  : RLEVFCEDRLGLTRELLDLLVLRGIDLRGIEIDPI--------GRIYLNFAELEFESFSSLMAE-IRRIAG--VTDVRTVPW

GlnD_Mtu_1707969        730-805  : ILEVRAMDRAGLLALLAGALEGAGAGIVWAKVNT-----FGSTAADVFCVTVPAELDARAAVEQHLLEVLGA-SVDVVVDEP
RelA_Eco_1710082        668-743  : VVRVVANDRSGLLRDITTILANEKVNVLGVASRSD---TKQQLATIDMTIEIYNLQVLGRVLGK-LNQVP--DVIDARRLHG
consensus/80%                      .a.h...s.sGhh.pa..hhsp.shsa..h............s...h..................b...............
```

```
Secondary structure                .EEEEEEEEE.......HHHHHHHHHH...EEEEEEE.....EEEEEEE..HHHHHHHHHH.......EEEEEEEE..
                                                                      #     #         #           #          ###
LrpA_Pfu_11278083        63-137  : SLVTITGVDTK---PEKLFEVAEKLKEYDFVKELYLSSGDHMIMAVIWAKDGEDLAEIISNKIGKIEGVTKVCPAIIL
Lrp_Eco_126474           73-150  : SLLVFVEITLNRGAPDVFEQFNTAVQKLEEIQECHLVSGDFDYLLKTRVPDMSAYRKLLGETLLRLPGVNDTRTYVVM
AsnC_Eco_114264          66-142  : DVGCFIGIILKS--AKDYPSALAKLESLDEVTEAYYTTGHYSIFIKVMCRSIDALQHVLINKIQTIDEIQSTETLIVL
SSO0157_Sso_13813287     64-137  : ENEIRAIVMVQSTPQIPTPEISKKIAKIPGVEVVYETTGDYDILVIVRGTNITSINRTIDE-IRSIQGVVGTNSTIIL
Lrp_Pmi_1922905          66-143  : SLLVFVEITLNRGAPDVFEQFNMAVQKLEEIQECHLVSGDFDYLLKTRVPDMSAYRKLLGETLLRLPGVNDTRTYVVM
Lrp_Pmu_12720483         70-147  : PLLVIVEITLIRGKPDVFEEFNAAVQQLDEIQECHLVSGDYDYLLKTRVADMAAYRKLLGTTLLRLPGVNDTRTYVVM
CC0805_Ccr_13422052      67-144  : ALLIFVEVVLERTTGDVFEEFGAAVRRAPEILECHMVAGGYDYLIKARVRDMEAYRSFLGDILVKMPGVRETRTYAVL
PutR_Atu_7388033         71-148  : DHIAFAEVKLSDTREKALEEFNTAVRKIKEVEECHMIAGAYDYLLKVRTSDIRKYRRVLGEKISSLPSVSNTSTFVVM
BkrD_Ppu_1168677         65-141  : DVNVFIHVSLEKQVEQSLHRFEEEIAERPEVMECYLMTGDPDYLLRVLLPSIQALERFLDY-LTRLPGVANIRSSFAL

PHS023_Pho_7430134         1-75  : MVTAFILMVTA---AGKEREVMEKLLAMPEVKEAYVVYGEYDLIVKVETDTLKDLDQFITEKIRKMPEIQMTSTMIAI
SSO5522_Sso_13813398       1-83  : MASAIVLINTD---AGGEEEVFEKLKNMNEVVEAYVVYGVYDIVIKVETENMDKLKNFISNSIRKIPKVRSTLTMIIM
PAB7115_Pab_7521755        1-76  : -MDVFILLVIR---PGLENEVYEKLKNRPEVKEIYKVYGDYDIVLRLSIDGIKALDKFHDEVLRKLPGIELSETLIAS
TVN0295_Tvu_13541126       1-77  : MSIAFVLISTV---PGKEHEVYNKVLKINGVVEAHPLFGEYDIIVKLDMKDFTEIGEVVIEKIRTIDGVIDTKTLTGI
AF1404_Afu_2649171         1-77  : MALGFVLIKVS---PGKERKVYDALAALDEVEELYPLFGEYDLIAKVVVRSFEELSDVVVNKIRSIDGVIETKTLTGA
SC7A1.05c_Sco_7479730      1-77  : MVQAYILIQTE---VGKASTVAETISKLPGVIQAEDVTGPYDVIVRAQSDTVDELGRMVVAKVQQVDGITRTLTCPVV
SSO10340_Sso_13815706      4-77  : VVRAYILVSTT---VGKEMEVADMAKKVSGVIRADPVYGEYDVVVEVEAKSSDDLKKVIYE-IRRNPNIIRTVTLIVM
msl0718_Mlo_14021642       1-78  : MKAFFVQIKCE---LGKSYEVASALADAEIASEIYSTAGNYDLLAKFYVDDEEDVGHFVNEKVQILPGIKDTFTVVTF

ORF2.N_Sno_8050857       70-148  : SGQGCLAFVEIDCAPGRLLDVARALAAAPHVVALSHVTGDRDLQLNVMARDPAMLSRWVTHDLAALDGVRAARTHLAG
ORF2.C_Sno_8050857      233-310  : SEWPVTVSFWGQVPAARLREVTRRVTGMREVRLCASVISRHNLHLVAWVRSLDDAQRFEVRLAERAADLTVTERAVAL
SC4B5.04.N_Sco_4007672   66-141  : LGQDQWTLRLRC-VPDSSAAIADALAKRPDTNWIGVASGGTEVILGTRPRSPGARDDLLLGKLPRTPRVLEIHAHQML
SC4B5.04.C_Sco_4007672  225-301  : LGYRTDAALWITTSPAALHRVGEALAAHDEIAHVAATAGPSNIIATAVVRSTADLYTYLSGPLGHLEGVQHLEASPLL

IPMS_Pae_*              361-438  : NRHIEAYVWLKVANNVYTTSVARRVATLKNVVSVSEITGDYDIVVKIIAENTEELNRVIEN-IREIKGVASTLTSIVL
LeuA-2_Sso_13814162     381-457  : PEQIEALISVKCDSNVYTTSVTRRLSVINGVKEVMEISGDYDILVKVQAKDSNELNQIIES-IRATKGVRSTLTSLVL
LeuA-2_Sto_15623321     381-457  : PENIEALISVKCESNVYTTAVTRRLSVIPGVKEVMEISGDYDILVKVEAKDPNELNQIIEN-IRAVKGVSSTLTSLVL
consensus/80%                      .......h............h...a.....h..h....Gpbphha.h...s.....ph....a..h..a.......h.
```

Figure 6.2 (a) Alignment of the most diverse members of the ACT domain. The secondary structure assignment that is indicated above the alignment was derived from the crystal structure of the *E. coli* 3-phosphoglycerate dehydrogenase (SerA; PDB code 1PSD). The amino acid residues that are involved in the binding of the effector serine in SerA, are indicated above the alignment (x). In addition, the ligand response mutations that were determined for the *E. coli* small subunit of the acetolactate synthase (IlvH) and aspartokinase (LysC) are indicated with * and $, respectively. The 80% consensus shown below the alignments was obtained as described above, and the position numbers on the left side indicate the limits of the domains. Also a structural alignment of the two regulatory ACT-like domains of the *E. coli* threonine deaminase (THD1) is included, with a mapped ligand response mutation (+), indicating the divergence on the sequence level between these domains and the genuine ACT domains. Secondary structure from the first repeat of THD1 is indicated above the alignment. The 80% consensus shown below the alignments was obtained using the following amino acid classes (4): small (*s*, ACGSTDNVP; *shaded yellow*), polar (*p*, YWHKREQDNST; *shaded green*), big (*b*, FILMVWYKREQ; *gray shading*), and hydrophobic (*h*, ILVCAGMFYWHTP; *shaded black*). Of the hydrophobic residues, the aliphatic subset (h, ILVA) is in black with red shading. The position numbers on the *left side* indicate the limits of the domains. **(b)** Multiple alignment of RAM domains. By using the structure of *P. furiosus* LrpA[16], the secondary structure of the RAM domain has been deduced, as is indicated *above* the alignment (*E*, β-strand; *H*, α-helix). The sequences are grouped by domain architecture as follows: 1, Lrp-like transcriptional regulators and homologs (HTH-RAM); 2, stand-alone versions of the RAM domain; 3, duplication of Lrp-like transcriptional regulator; and 4, RAM domain present in crenarchaeal 2-isopropylmalate synthase. The ligand response mutations as determined for the *E. coli* Lrp by Platko and Calvo[17] are indicated abovethe alignment (#). All protein sequences were obtained from ENTREZ, and GenBank™ identifiers are indicated for each protein. Species abbreviations are as follows: Aea: *Aquifex aeoliticus*; Afu: *Archaeoglobus fulgidus*; Atu: *Agrobacterium tumefaciens*; Bsu: *Bacillus subtilis*; Ccr: *Caulobacter crescentus*; Eco: *E. coli*; Mlo: *Mesorhizobium loti*; Mtu: *Mycobacterium tuberculosis*; Pab: *Pyrococcus abyssi*; Pfu: *P. furiosus*; Pho: *Pyrococcus horikoshii*; Pmi: *Proteus mirabilis*; Pmu: *Pasteurella multocida*; Ppu: *Pseudomonas putida*; Sco: *Streptomyces coelicolor*; Sno: *Streptomyces noursei*; Sso: *S. solfataricus*; Sto: *S. tokodaii* and Tvu: *Thermoplasma vulcanium*.

Figure 7.2

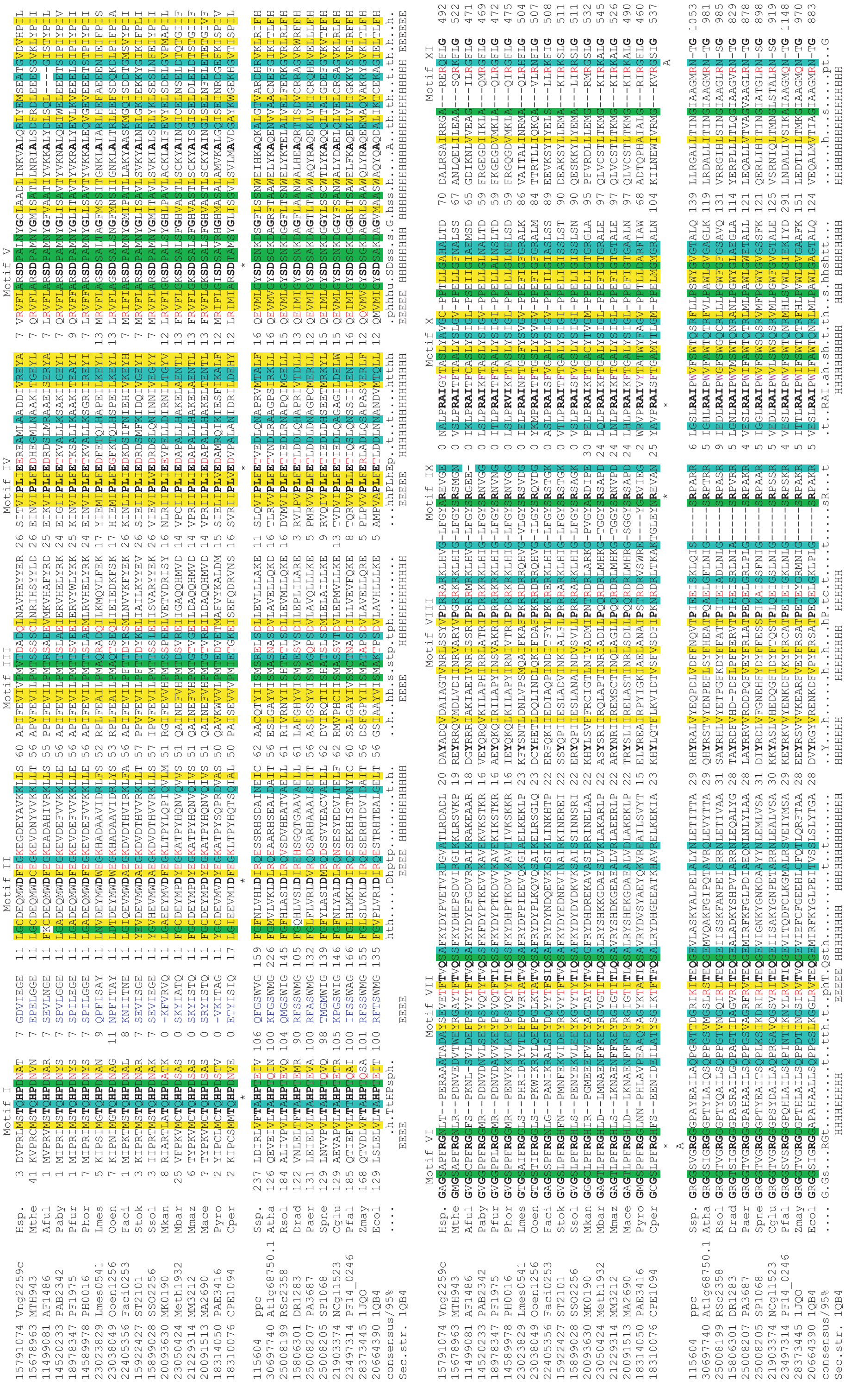
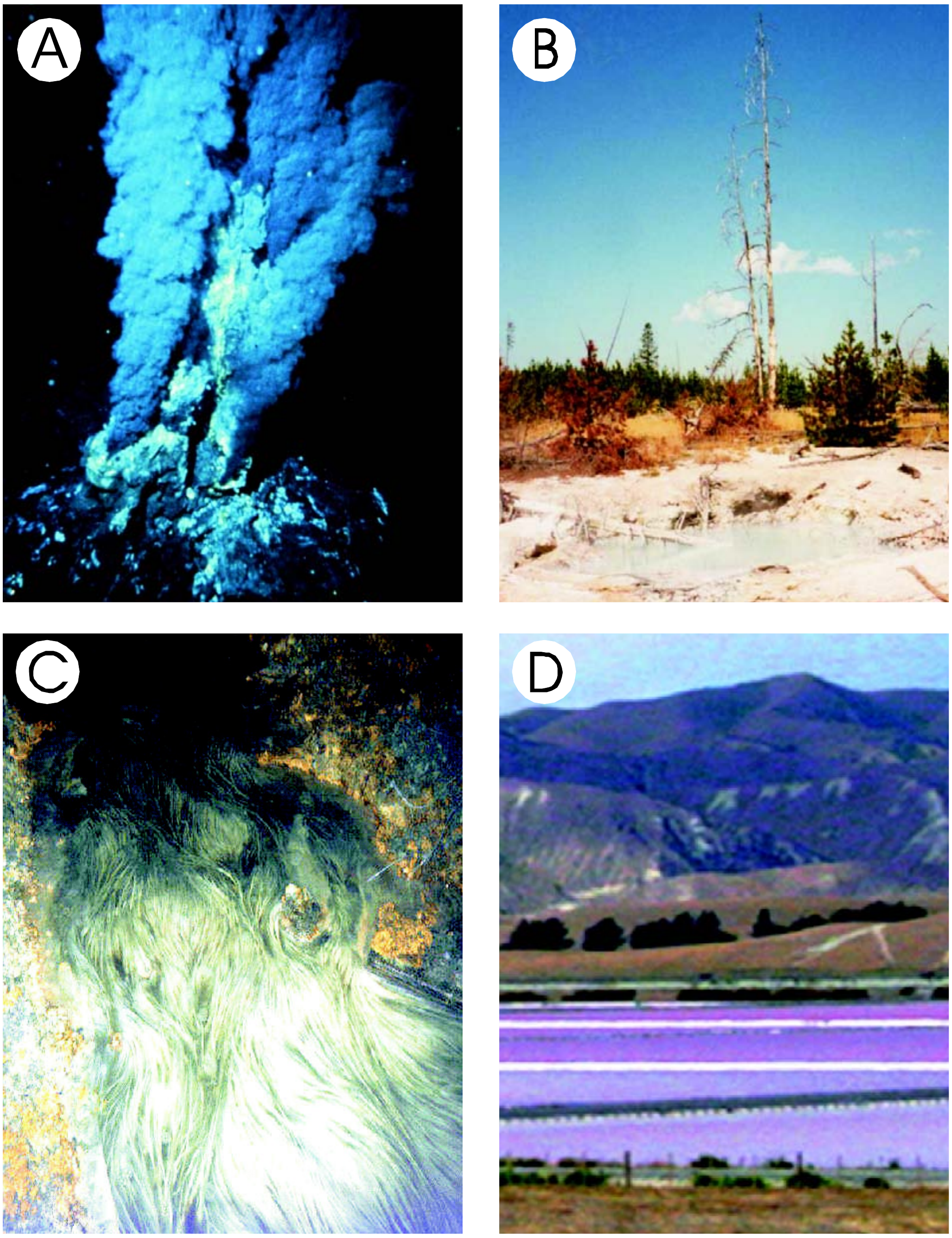


Figure 7.2 Multiple alignment of the conserved core of archaeal, bacterial and eukaryote PEPC sequences. The sequences are denoted by Gene Identification (gi) numbers from the GenBank database, species abbreviation (see legend to Figure 7.1) and systematic gene numbers; proteins with available structure denoted by their PDB code. The positions of the first and the last residue of the aligned region in the corresponding protein are indicated for each sequence. The numbers within the alignment represent poorly conserved inserts that are not shown. Amino acids residues that are involved in the formation of the active centre are shown by asterisks; those that have an anticipated involvement in L-aspartate binding are denoted as "A" in the line between atPEPC and BE-PEPC families. Positions with identical amino acids in both families are boldfaced. The coloring is based on the consensus (calculated for all sequences in the alignment) shown underneath the alignment; *h* indicates hydrophobic residues (*ACFILMVWYH*), *t* indicates turn-forming residues (*ASTDNVGPENRK*), *p* indicates charged residues (*STEDKRNQH*), *s* indicates small residues (*AGSVC*), *a* indicates aromatic residues (*FYW*). The secondary structure elements correspond to those experimentally identified for 1QB4[15]. *H* indicates α-helix; *E*, β-strand. Sequence region of the β-strand 2 for 1QB4 and predicted by JPRED program[18] for gi|18310076 (CPE1094) as a query β-strand 2 for atPEPC are shown in blue.

Figure 10.2



Figuur 10.2 Extreme milieu's waar sommige archaea voorkomen: **(a)** *Mid-atlantic ridge, Atlantische oceaan* – Deze donkere pluimen van hete thermale vloeistof zijn het gevolg van onderaardse vulkanische activiteit en rijk aan opgeloste metalen. Als deze opgeloste metalen in contact komen met het koude zeewater, wordt een zwarte neerslag gevormd. Deze neerslag zinkt naar de zeebodem en creëert zo een schoorsteenachtige structuur, de 'black smoker'. De thermale vloeistoffen zijn rijk aan nutriënten, wat bepaalde organismen in staat stelt om in deze extreme hitte te leven, de (hyper)*thermofielen* [Foto: Peter Rona, NOAA]. **(b)** *Ragged Hills, Yellowstone National Park, USA* – Dit thermale gebied in Yellowstone NP is nog relatief jong en is ontstaan als gevolg van recente thermale activiteit. Door de stijging van temperatuur en zuurtegraad sterven de bomen en planten af en gaan hotsprings het landschap domineren. Deze hotsprings zijn vaak kokend heet en enorm zuur (pH < 3). In dit type hotsprings leven acidofielen als *Sulfolobus* en *Thermoplasma* [Foto: Thijs Ettema]. **(c)** *Iron Mountain, California, USA* – Waterstroom afkomstig uit een oude ijzer/koper mijn, die zich het beste laat vergelijken met accuzuur: extreem zuur (pH tussen 1.3 and 2.2) en extreem hoge metaal concentraties (tot 120 gram per liter). In deze extreme condities leven metallofielen als Ferroplasma, die energie winnen door metaalsulfides te oxideren [Foto: Katrina Edwards, WHOI]. **(d)** *Blenheim, Nieuw Zeeland* – Het zoutmeer op deze foto wordt gebruikt voor commerciële zoutwinning door middel van verdamping van zeewater, waardoor het zout zijn verzadigingspunt (30%) bereikt en uitkristalliseert. Deze extreem zoute meren, waar de zout concentratie 10 keer zo hoog is als in de zee, zijn vaak paars of roze van kleur. Deze kleur wordt veroorzaakt door micro-organismen die in staat zijn deze extreme zoute condities te weerstaan, de halofielen. De meest extreme halofielen behoren tot de archaea [Foto: Thijs Ettema].

CURRICULUM VITAE

Thijs Johannes Gerardus Ettema werd op 24 maart 1977 te Veghel geboren. Hij groeide op in Uden, waar hij de kleuterschool en basisschool doorliep. Op een flinke steenworp afstand van zijn ouderlijk huis genoot hij vervolgens middelbaar onderwijs aan het Kruisheren Kollege, waar hij in 1995 zijn VWO diploma behaalde.

Na als kind altijd al geïnteresseerd te zijn geweest in 'alles wat groeit en bloeit', besloot Thijs naar Wageningen te verhuizen, om daar de studie Biologie te volgen aan de aldaar gevestigde Wageningen Universiteit. Tijdens zijn specialisatie in de moleculaire en cellulaire biologie kwam hij in aanraking met het Laboratorium voor Microbiologie, waar hij besloot om een aantal afstudeerprojecten en stages te doen. Allereerst werkte hij aan een tweetal projecten waarin het suikermetabolisme van *Pyrococcus furiosus*, een hitte-minnend micro-organisme, centraal stond. Daarna volbracht hij meerdere afstudeerstages bij verschillende instellingen. Bij Plant Research International (PRI, Wageningen) werkte hij aan de genetisch modificatie van suikerbieten, met als doel de suikeropbrengst te verhogen. Vervolgens deed hij bij DSM-Gist (Delft) onderzoek aan de optimalisatie van heterologe genexpressie in het industriële modelorganisme *Aspergillus niger*. Tot slot specialiseerde hij zich in de bioinformatica gedurende een stage bij de Biocomputing Unit van het EMBL (Heidelberg, Duitsland).

Eenmaal terug in Nederland, studeerde Thijs *cum laude* af in september 2000, waarna hij per direct als promotieonderzoeker werd aangesteld bij het Laboratorium voor Microbiologie in de Bacteriële Genetica groep van prof. dr. John van der Oost. Hier heeft hij met behulp van vergelijkende en functionele genoomanalyse inzicht gekregen in diverse aspecten van de intrigerende biologie van de archaea. De resultaten van dit promotieonderzoek hebben geresulteerd in een aantal wetenschappelijke artikelen, welke gebundeld zijn in dit proefschrift.

Sinds oktober 2004 is Thijs werkzaam als Postdoc onderzoeker in de Computational Genomics groep van prof. dr. Martijn Huynen aan het CMBI (Radboud Universiteit, Nijmegen), waar hij onderzoek doet aan genoomevolutie en voorspelling van gen– en eiwitfuncties.

LIST OF PUBLICATIONS

Ettema, T.J.G., 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

Verhees, C.H., Koot, D.G., **Ettema, T.J.G.**, 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

Ettema, T.J.G., 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

Ettema, T.J.G., Huynen, M.A., de Vos, W.M., and van der Oost, J. (2003) TRASH: a novel metal-binding domain predicted to be involved in heavy-metal sensing, trafficking and resistance. *Trends Biochem Sci* 28, 170-173

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

Van Lieshout, J.F.T., Verhees, C.H., **Ettema, T.J.G.**, Van der Sar, S., Imamura, H., Matsuzawa, H., Van der Oost, J., and De Vos, W.M. (2003) Identification and molecular characterization of a novel type of alpha-galactosidase from *Pyrococcus furiosus*. *Biocatal Biotransf* 21, 243-252

Ettema, T.J.G., 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 phospho*enol*pyruvate carboxylase, a missing link in archaeal central carbohydrate metabolism. *J Bacteriol* 186, 7754-7762

Ahmed, H.*, **Ettema, T.J.G.***, 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

* Both authors contributed equally

Kaper, T., Talik, B., **Ettema, T.J.G.**, Bos, H., Van der Maarel, M.J.E.C., and Dijkhuizen, L. (2005) Extreme thermostable amylomaltase from *Pyrobaculum aerophilum* IM2 produces thermoreversible starch gels. *Appl Environ Microbiol* 71, 5098-5106

Brouns, S.J.J., **Ettema, T.J.G.**, Stedman, K.M., Walther, J., Smidt, H., Snijders, A.P.L., Young, M., Bernander, R., Wright, P.C., Siebers, B., and Van der Oost, J. (2005) *The hyperthermophilic archaeon Sulfolobus - from exploration to exploitation*. Young, M.a.I., B. Geothermal Biology and Geochemistry in Yellowstone National Park, in press

Ettema, T.J.G., de Vos, W.M., and Van der Oost, J. (2005) Discovering novel biology by *in silico* archaeology. *Nat Rev Microbiol*, in press

Ettema, T.J.G., Brinkman, A.B., Lamers, P.P., Kornet, N.G., van der Oost, J., and de Vos, W.M. (2005) A novel type of copper-responsive transcription regulator in archaea, submitted for publication

Ettema, T.J.G., Ahmed, H., Siebers, B., and Van der Oost, J. (2005) The non-phosphorylating glyceraldehyde-3 phosphate dehydrogenase in *Sulfolobus solfataricus*: a key-enzyme of the semi-phosphorylative Entner-Doudoroff pathway, submitted for publication