Extremophiles (1998) 2:305–312

REVIEW

**Christoph W. Sensen · Robert L. Charlebois
Cynthia Chow · Ib Groth Clausen · Bruce Curtis
W. Ford Doolittle · Michel Duguet · Gael Erauso
Terry Gaasterland · Roger A. Garrett · Paul Gordon
Ineke Heikamp de Jong · Alex C. Jeffries
Catherine Kozera · Nadine Medina · Anick De Moors
John van der Oost · Hien Phan · Mark A. Ragan
Margaret E. Schenk · Qunxin She · Rama K. Singh
Niels Tolstrup**

Completing the sequence of the *Sulfolobus solfataricus* P2 genome

Received: January 22, 1998 / Accepted: February 16, 1998

Abstract The *Sulfolobus solfataricus* P2 genome collaborators are poised to sequence the entire 3-Mbp genome of this crenarchaeote archaeon. About 80% of the genome has been sequenced to date, with the rest of the sequence being assembled fast. In this publication we introduce the genomic sequencing and automated analysis strategy and present intial data derived from the sequence analysis. After an overview of the general sequence features, metabolic pathway studies are explained, using sugar metabolism as an example. The paper closes with an overview of repetitive elements in *S. solfataricus*.

Key words *Sulfolobus solfataricus* · Archaea · Genomic sequence · MAGPIE

Communicated by J. Wiegel

C.W. Sensen (✉) · P. Gordon · A.C. Jeffries · M.A. Ragan · R.K. Singh
National Research Council of Canada, Institute for Marine Biosciences, 1411 Oxford Street, Halifax, NS, Canada B3H 3Z1
Tel. +1-902-426-7310; Fax +1-902-426-9413
e-mail: sensencw@niji.imb.nrc.ca

C.W. Sensen · R.L. Charlebois · W.F. Doolittle · T. Gaasterland M.A. Ragan
Canadian Institute for Advanced Research, Evolutionary Biology Program, Canada

R.L. Charlebois · A. De Moors
University of Ottawa, Department of Biology, 30 Marie Curie, Ottawa, Ontario, Canada K1N 6N5

C. Chow · B. Curtis · W.F. Doolittle · C. Kozera · M.E. Schenk
Dalhousie University, School of Medicine, Department of Biochemistry, Sir Charles Tupper Building, Halifax, NS, Canada B3H 4H7

I.G. Clausen · N. Tolstrup
Novo Nordisk A/S, Screening Biotechnology Enzyme Research, Novo Alle Ibmi, DK-2880, Bagsvaerd, Denmark

M. Duguet · N. Medina
Université Paris Sud, Institut de Genetique et Microbiologie Laboratoire d'Enzymologie des Acides Nucleiques, 15, Rue Georges Clemenceau Batiment 400, FR-91405 Orsay, Cedex, France

R.A. Garrett · H. Phan · Q. She
University of Copenhagen, Institute of Molecular Biology, Sølvgade 83H, DK-1307, Copenhagen K, Denmark

T. Gaasterland
Argonne National Laboratory, Mathematics and Computer Science Division, 9700 S. Cass Ave, Argonne, IL 60439, USA; and University of Chicago, Department of Computer Science, Ryerson Hall, 100 E. 58th St, Chicago, IL 60637, USA

G. Erauso · I. Heikamp de Jong · J. van der Oost
Wageningen Agricultural University, Laboratory of Microbiology, Hesselink van Suchtelenweg 4, NL-6703 CT, Wageningen, The Netherlands

Introduction

The *Sulfolobus solfataricus* P2 genome project started four years ago as the first complete genome sequencing project in Canada. It was also the first archaeal genome project funded worldwide. The *S. solfataricus* project was initiated by members of the Evolutionary Biology Program of the Canadian Institute for Advanced Research, who had (and still have) strong interests in phylogenetic and evolutionary issues involving archaea.

The reasons for sequencing this 3-mega-base pair (Mbp) crenarchaeote genome were manifold. The *S. solfataricus* genome is among the largest genomes known for Archaea. Its low G+C% content (36%) renders it relatively easy to sequence. Several plasmids and viruses from Sulfolobales have been characterized and sequenced, and progress has been made in developing shuttle vectors for *Sulfolobus* species. *S. solfataricus* can be grown aerobically in large quantities at temperatures between 65° and 75°C and at low pH values (2–5), thus many of its genes could be of commercial interest.

After starting in Canada with a single automated sequencer in 1994, the project is now organized as an international collaboration. In 1995, a member of Argonne National Laboratory (USA), Terry Gaasterland, joined the Canadian *S. solfataricus* researchers to collaborate on automated genome analysis and annotation. At the end of 1996, three European laboratories (from Denmark, France, and The Netherlands), and Novo Nordisk (Denmark) as an in-

dustrial partner, joined the Canadian team with the mandate to contribute to finishing the *S. solfataricus* genome by sequencing one-third of the genome. Here, we present the organization of the project and data that have derived from the analysis of the *S. solfataricus* genome.

Genomic cloning and mapping strategy

Genomic DNA was isolated from cells which were grown for 1 week in 2-l shaker flasks. The 3-Mbp *S. solfataricus* P2 genome (DSM #1617; ATCC #35092) was partially digested with *Bam*HI or *Hind*III and subcloned into the cosmid vector Tropist3 (De Smet et al. 1993). The average insert size in the cosmid libraries that were created using this vector was 40 kilobase pairs (kbp). A number of cosmids equivalent to a 14-fold coverage of the genome were studied using the landmark strategy (Charlebois et al. 1991) and Southern hybridization of cosmids to cosmids. Cosmids covering about 70% of the *S. solfataricus* genome and building clusters between 40 and 200 kbp were identified by these two methods.

The ends of about 400 phage lambda (λ) clones, equivalent to a threefold coverage of the genome, were sequenced, using T3 and T7 primers, to complement the cosmid libraries. These λ clones were mapped onto the existing cosmid sequences through BLASTN homology searches (mapping by sequencing). Cosmid contigs and λ contigs are extended through several iterations of this process. Contig ends that cannot be extended based on the 800 T3 and T7 sequences will be hybridized against the λ libraries to find extensions. In the near future, we expect five to ten remaining gaps in the *S. solfataricus* genome. The closure of these gaps will be done using long PCR fragments or similar strategies. Figure 1 illustrates the cloning strategy for the *S. solfataricus* genome.

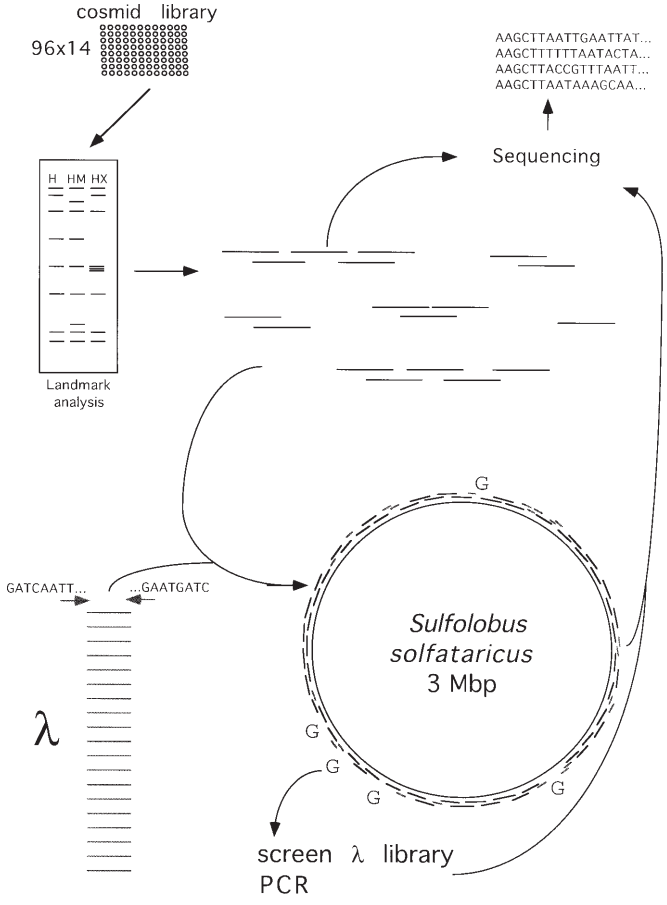

Fig. 1. Cloning and mapping strategy. G-gap

Sequencing strategy, sequence assembly, and primer calculation

Sequencing is performed in four locations: Halifax, Copenhagen, Paris, and Wageningen. In total, the project runs three ABI 373A stretchliner, two LI-COR 4200 IR^2, one LI-COR 4000L, and two Amersham Vistra automated sequencers. The data flow between the laboratories is handled via the Internet through ftp connections.

For sequencing, each cosmid or λ clone chosen is fragmented by nebulization, followed by Klenow fill-in, and subcloned into the double-stranded plasmid vector pUC18 using blunt-end ligation. The average insert size of the random clones is 1–2 kbp. Before sequencing, the clones are size-selected on agarose gels for an insert size of 1–2 kbp. Rigid naming conventions for all clones and subclones are enforced to ensure that all sequencing reactions are assembled into the right group of contigs.

During the compilation of sequence data associated with a particular cosmid or λ clone, the sequence is assigned to one of four possible states. The first state is the primary sequencing state where random pUC18 clones that were created using nebulized DNA are sequenced and assembled without error correction. Primary sequencing of random pUC18 clones is performed on the single-dye, four-lane machines (i.e., all machines except the ABI 373A sequencers) to an average coverage of three sequence readouts per base. This results in a few gaps remaining in the contig. In this phase, for a 40-kbp insert, typically 60–70 clones are sequenced from both ends with the LI-COR machines. This step requires about three times as many sequences on the Amersham Vistra machines because of their shorter read lengths. The LI-COR 4200 IR^2 machines are the best sequencers for this sequencing phase, as they are capable of sequencing from both ends of the pUC18 clone in the same sequencing reaction. This saves 50% of the work and consumable items in the primary sequencing state. Compared to ABI373A or 377 machines, the cost savings by LI-COR 4200 IR^2 two-dye sequencing are on the order of 80% because of the longer readout length (average, 850 bp per read).

The second state is the linking state where, to link the assembled primary sequencing contigs into one contig, walking primers are designed at the ends of all contigs that contain at least three sequencing readouts. All sequencing

primer calculations are performed using the program "Osprey" (Gordon et al., in manuscript). This program can calculate walking primers for the ends of contigs as well as polishing primers to disambiguate contigs (see polishing state). The gaps are closed by directed primer-walking reactions on the cosmid or λ DNA templates. All primer-walking reactions are performed on the ABI 373A automated sequencing machines using a sequencing protocol proprietary to the National Research Council of Canada. Typically, for a 40-kbp insert, 30–40 walking primers need to be applied to obtain a single contig.

The third state is the polishing state. In this state, the last ambiguities in the linked contig are resolved and all of the sequence becomes double stranded (sequenced on both DNA strands). The same sequencing technology that was used in the linking state is used for polishing primers. Typically, 60–80 polishing primers are used to finish a 40-kbp contig.

The last state is the finished state. All of the contig is double stranded and no ambiguities remain. Sequence assemblies are tested by comparing a computer-generated "virtual" agarose gel with digests of the cosmid or λ clone, using at least three different restriction enzyme patterns. After the confirmation of the sequence assembly, thorough sequence analysis and annotation are performed.

Our strategy has several advantages for collaborative genome projects. Relatively large contigs (25–45 kbp) are generated from the start of the project, allowing early access to publishable data. Repeats can be separated into individual contigs and handled separately, which is important for highly repetitive genomes. The management overhead is low, because all primary clones and all walking primers can be discarded after a contig is finished. All that needs to be preserved for the future in the *S. solfataricus* example are 60–70 cosmid clones and approximately the same number of λ clones.

Sequence assembly is performed using the Staden package (Bonfield et al. 1995) on the UNIX machines of the Canadian Bioinformatics Resource.[1] The Staden package generates a link between the sequence assembly and the original trace data, which allows the *S. solfataricus* team to exchange only the trace data files, which were generated on the automated sequencing machines, without loss of information at any site. The performance of the *S. solfataricus* project is shown in Fig. 2. The error rate in the *S. solfataricus* project, as measured by the comparison of overlapping finished contigs, is of the order of 1 error in 5000–10 000 bp.

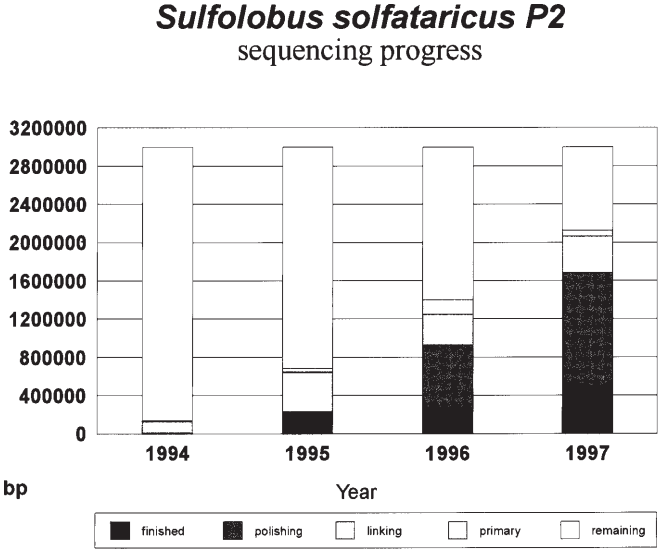

Fig. 2. Sequencing progress in the *Sulfolobus solfataricus* P2 genome project

Automated sequence analysis and annotation

The sequence produced in the *S. solfataricus* project is constantly analyzed using the automated genome analysis system, MAGPIE (Multipurpose Automated Genome Project Investigation Environment) (Gaasterland and Sensen 1996a,b). MAGPIE was developed mainly using the *S. solfataricus* data, but is now used in several other genome analysis projects as well. Figure 3 shows the data flow in the *S. solfataricus* genome analysis. An example of the MAGPIE analysis of the *S. solfataricus* data can be accessed at http://niji.imb.nrc.ca/sulfolobus/. The *S. solfataricus* team is using an entire suite of analysis tools (Table 1).

MAGPIE can be customized to reflect the researchers' needs for the analysis of individual contigs. In addition, the sequence analysis needs to change with the state of the sequence. In the primary sequencing state, the analysis is restricted to the identification of contaminant clones (i.e., *E. coli* genomic DNA or vector sequences). In the linking phase, the first functional assignments can be performed on the emerging open reading frames (ORFs), but the assignment of function to ORFs is mainly carried out on the polishing and finished state sequences. The minimal ORF size in the *S. solfataricus* genome project has been set to 100 amino acids. Once a contig is completely finished, intergenic features such as promoters and terminators are identified in addition to the functional assignments of ORFs (Gordon et al., unpublished). All features are presented in graphical overviews of the genomic sequence (Gordon et al., unpublished).

Automated functional assignments can be modified by human intervention and annotations can be added. All corrections and final assignments are stored in local databases

[1] The Canadian Bioinformatics Resource (CBR-RBC) is a new initiative of the National Research Council of Canada to bring high-performance bioinformatics to Canadian researchers and Canadian industry. Currently, it consists of a distributed network of high-end UNIX machines, located at five NRC institutes. Many of the original CBR-RBC concepts were developed in the context of the *Sulfolobus* project.

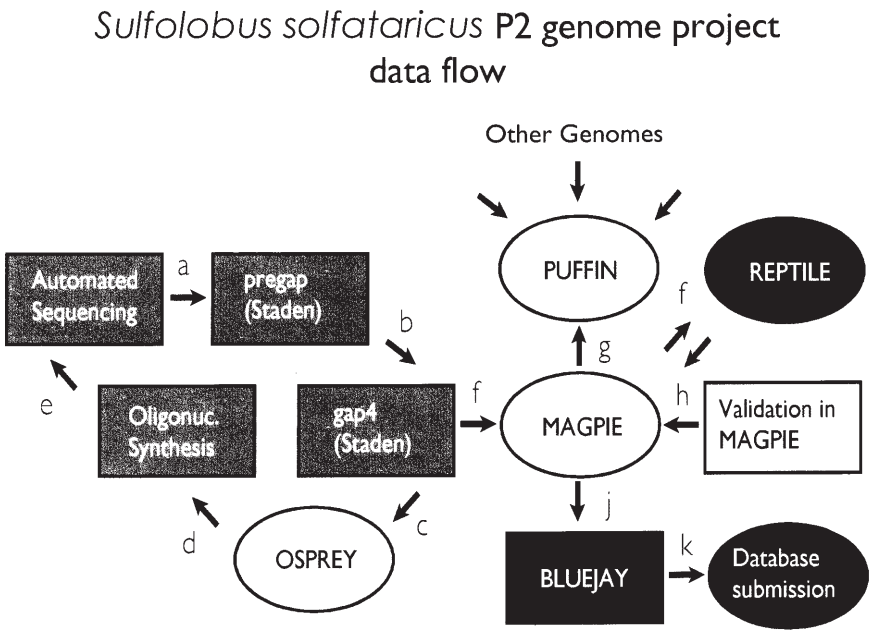

Fig. 3. Data flow in the *Sulfolobus solfataricus* P2 genome project. *a*, trace data, ABI or SCF; *b*, SCF data and experimental data files; *c*, FASTA formatted files and sequence assembly reports; *d*, GenBank formatted files; *f*, FASTA formatted files, sequence assembly reports, multiple alignment overviews; *g*, FASTA formatted genomes, sequence annotations; *h*, user input; *j*, MAGPIE databases and user annotations; *k*, database submission form

Table 1. Softward tools used to analyze the *Sulfolobus solfataricus* genome

Tool	Location	Target	Tool type	State			
				Pri	Lin	Pol	Fin
BLAIZE	blitz@ebi.ac.uk	Protein	Protein similarity		X	X	X
BLASTN	blast@ncbi.nlm.nih.gov	DNA	DNA similarity	X	X	X	X
BLASTP	blast@ncbi.nlm.nih.gov	Protein	Protein similarity			X	X
BLASTX	blast@ncbi.nlm.nih.gov	DNA	Protein similarity	X	X		
BLOCKS	blocks@howard.fhcrc.org	Protein	Motif search tool			X	X
FASTAE	fasta@ebi.ac.uk	DNA	*E. coli* similarity search	X			
FASTAN	fasta@ebi.ac.uk	DNA	DNA similarity	X	X	X	X
FASTAP	fasta@ebi.ac.uk	Protein	Protein similarity			X	X
FASTAV	fasta@ebi.ac.uk	DNA	Vector similarity search	X			
FASTAW	fasta@ebi.ac.uk	DNA	DNA similarity update	X	X	X	X
PREPRO	predictprotein @embl-heidelberg.de	Protein	Structure prediction			X	X
PROSEARCH	Local	Protein	Motif search tool			X	X
PROTERM	Local	DNA	Promoter and terminator		X	X	X
SAPS	Local	Protein	Statistical analysis			X	X
SHNDEL	Local	DNA	Shine Dalgarno identification		X	X	X
SPUTNIK	Local	DNA	Motif search tool			X	X
TBLASTN	blast@ncbi.nlm.nih.gov	DNA	DNA similarity		X	X	X
TBLASTX	blast@ncbi.nlm.nih.gov	DNA	Protein similarity		X	X	X
TRNASCAN	Local[6]	DNA	DNA similarity			X	X

that are used as the template for automated database submission. An important notion in the automated analysis is the level of confidence that is assigned to individual assignments (Gaasterland et al. 1994). Level 1 confidence means that the ORF is homologous to some other gene in the databases with very high (90%–100%) confidence. Level 2 confidence means that in 80% of all cases, the automated assignment is correct, and the remaining genes are similar, but not homologuos. Level 3 confidence means that there are weak sequence similarities to other genes in the databases, but the overall similarity of the ORF to the database entries is too low to make a confident assignment. Level 4 evidence is discarded because it can be considered to be background noise that is still reported by the tool server at the lower cutoff level. As much as 80% of the toolserver output is discarded on arrival in the MAGPIE space. The data collection and analysis are rerun at regular time intervals, to reflect the changes in the public databases in the analysis of the *S. solfataricus* genome. The UNIX servers and workstations of the Canadian Bioinformatics Resource allow a complete analysis of the *S. solfataricus* genome to be repeated every 24 h.

All the *S. solfataricus* ORFs are classified into functional categories (wherever possible) and subsequently compared to a complete set of ORFs from all other completed genomes using the PUFFIN software (Gaasterland and Ragan, in manuscript). In the process, sets of genes and gene families are identified using BLAST and FASTA similarity searches. Based on the similarity search results, statistical data are produced that allow the comparison of gene content between organisms.

General sequence features

The *S. solfataricus* P2 genome has an average G+C% content of 36%, which makes it relatively easy to identify ORFs using start and stop codons (*S. solfataricus* can use ATG, GTG, and TTG as start codons). *S. solfataricus* has approximately 1 ORF with a size of at least 100 amino acids per kbp, which will lead to a total number of about 3000 ORFs once the genome is complete. Presently, the average size of an ORF is 841 bp, and the average size of an intergenic spacer is 191 bp. These numbers might shift slightly toward larger intergenic regions and a smaller average ORF size when the true start codons for *S. solfataricus* genes have been identified (at present, the outermost start codon is used to determine the ORF boundaries). Only a small fraction of the ORFs overlap with ORFs on the opposite DNA strand. To date, we have identified very few ORFs that might overlap with an ORF on the same DNA strand (Sensen et al. 1996).

Intergenic regions show a sharp drop in the G+C% content, to approximately 25%. Some of the ORFs show a dramatically different G+C% from the average, typically between 50% and 60%. In all instances, these ORFs could be identified as insertion elements. Another exception to the generally low G+C% of the *S. solfataricus* genome is the 16S–23S rRNA region with an average G+C% of 60%. Coding ORFs typically show an A + G (purine) content of 55% on average (measured on the coding DNA strand), similar to results from all currently completed genomes, including eubacteria and yeast. The reasons for the purine enrichment on the coding strand are still poorly understood.

Functional assignments with level 1 confidence have been made for about 40% of the *S. solfataricus* ORFs. Adding level 2 confidence assignments raises the percentage to between 60% and 70%. This ratio of identified to nonidentified ORFs is very similar to that reported for other sequenced archaeal genomes. In eubacterial genomes, usually a higher percentage of ORFs can be associated with a particular function because more research has been done to isolate and characterize particular genes. Thus, functional assignment is an area of future work for the archaeal research community. Comparisons between the genomes of the crenarchaeote *S. solfataricus* and the euryarchaeote *Methanococcus jannaschii* (Bult et al. 1996) show a large number of genes that are unique to the *S. solfataricus* genome. This result could probably indicate that the larger *S. solfataricus* genome (about twice the size of the *Methanococcus* genome) has a greater number of genes. The complete picture will emerge only after the complete genome of *S. solfataricus* is available.

Many of the *S. solfataricus* genes are clustered according to their function (Charlebois et al. 1996). Although not verified experimentally (except in a few cases studied in the EU Extremophile program), many of the *S. solfataricus* genes appear to be organized in operons (for example, the *his* operon, Charlebois et al. 1997). The clustering of genes makes it possible to predict functions for ORFs that are associated with other ORFs of known function, but these predictions must be verified experimentally.

The only genes in the *S. solfataricus* genome in which introns have been identified to date are 6 (of 20 currently identified) tRNA genes. They lie within the tRNA genes for the anticodons ATG (Met), CCC (Pro), TAC (Tyr), TCG (Ser), TGG (Trp), and TTA (Leu). Two of these, ATG (Met) and TCG (Ser), were identified earlier (Kaine et al. 1983; Kaine 1987). This ratio of intron-containing to intron-lacking genes is similar to the results from other archaeal genome projects.

Phylogenetic analyses of *Sulfolobus solfataricus* ORFs

Initial phylogenetic analyses using ORFs identified with level 1 confidence do not result in a conclusive phylogeny. Detailed analyses will be presented elsewhere, but generally at least as many strong homologies exist between *S. solfataricus* ORFs and their eubacterial counterparts as between *S. solfataricus* and eukaryotic genes. Even when a single operon is studied (Charlebois et al. 1997), the phylogenetic trees derived from the analysis of individual ORFs yield very different reconstructions of phylogeny. The proteins involved in information processing seem to be the "most consistently eukaryotic" part of an archaeon, while most other aspects are more or less shared with the other phylogenetic domains. A large number of the *S. solfataricus* genes seem to be shared only with other Archaea, because no homologs in the other phlyogenetic domains were identified through database searches. At present, with only the *S. solfataricus* data available from the crenarchaeote branch of the phylogenetic tree, it is too early to predict how many of the *S. solfataricus* genes are exclusively crenarchaeal. This situation will change once the *Pyrobaculum aerophilum* genome is available. (For a complete list of genome projects, see: *http://www.mcs.anl.gov/home/gaasterl/genomes.html*).

Studying the metabolic pathways of *Sulfolobus solfataricus*

After all ORFs in a contig have functions assigned in the automated genome annotation process, further work is necessary to verify the predicted functions. This work, although tedious and time consuming, is essential to gain an

understanding as to how metabolic pathways have evolved and how they may function in vivo. Attempts to express certain *S. solfataricus* genes are being carried out in several of the laboratories collaborating in the *S. solfataricus* project. The following section on sugar metabolism in *S. solfataricus* provides a detailed example of the efforts being made to work on biochemical characterizations of certain metabolic pathways, once the sequence is generated.

Sugar metabolism

It has been reported that, in addition to peptides, *S. solfataricus* can use a number of sugars as carbon or energy sources, including glucose, xylose, sucrose, lactose, maltose, and rhamnose (De Rosa et al. 1975, 1984). Because no genome sequences of saccharolytic Archaea have yet been completed, we have examined the metabolic pathways involved specifically in archaeal sugar catabolism. *S. solfataricus* was originally isolated from a solfataric field at elevated temperatures (75°–90°C) and low pH values (1.0–3.0). Decomposing materials from higher plants are present under these conditions, and polymers including cellulose, hemicellulose (xylans), and starch are potential carbon sources. Although the rather extreme physical and chemical conditions of the *S. solfataricus* niche might enhance the chemical degradation of glycosidic bonds, the *S. solfataricus* genome contains at least two genes that potentially encode extracellular glycosyl hydrolases. Both these show significant similarity to the GH-C of the glycosyl hydrolase superfamily, consisting of β-1,4 specific xylanases and cellulases (families 11 and 12). Both the gene products exhibit an N-terminal hydrophobic (signal) peptide, suggesting they have an extracellular location. In addition, both genes are clustered with a gene encoding (intracellular) glycosyl hydrolase, which shows low but significant similarity to β-xylosidase from *Bacillus stearothermophilus* (family 52), in which the xylosidase gene is also adjacent to a xylanase-encoding gene (Baba et al. 1994). Although the substrate specificity of the *S. solfataricus* gene products remains to be identified, it is likely that these proteins are involved in the degradation of plant polysaccharides.

Another gene cluster of interest is a potential operon that encodes an α-amylase, a glycogen-debranching enzyme (pullulanase), and a maltooligosyl-trehalose synthase, sharing a high degree of sequence similarity with the products of the recently described *treZXY* gene cluster from *Sulfolobus acidocaldarius*, proposed to be involved in trehalose biosynthesis (Maruta et al. 1996). These gene products do not contain potential signal sequences, suggesting that their location is cytoplasmic. After the sugar polymers are degraded to cellobiose, xylobiose or maltose, specific transport systems are likely to be involved in the import of the oligo- and disaccharides across the cytoplasmic membrane. Two types of secondary transporters have been identified in the course of this project: high-affinity ATP-binding cassette (ABC) - transporters (14 copies to date), and several members of the Facilitator superfamily. To date, no phosphotransfer systems (PTS) have been found in *S. solfataricus* or in other Archaea.

One gene cluster probably encodes an ABC-type sugar transporter (highest similarity with the maltose import system), including an extracellular, periplasmic binding protein, two integral membrane subunits, and a cytoplasmic ATPase subunit. A second copy of this gene cluster is present in the *S. solfataricus* genome; however, it does not contain a copy of the gene that encodes the putative binding protein. This pattern also occurs in ABC-type transport systems for branched amino acids (ILV): one cluster occurs with, and one without, the gene encoding the binding protein. A similar observation has been made in the recently completed genome sequence of the heterotrophic archaeon *Archaeoglobus fulgidus* (Klenk et al. 1997).

The other type of transporter resembles the bacterial arabinose/H^+ symport system. Two copies of the latter type have been identified. In addition, another member of this Facilitator family, encoding a putative sugar uptake system, was isolated from *S. solfataricus* MT4, and found to be located adjacent to a gene encoding β-galactosidase family 1 (Prisco et al. 1995). These genes have also been found in the *S. solfataricus* P2 genome, but here they are separated by a transposon, ISC1439. In addition, two genes encoding α-glucosidase (family 31; eukaryal sucrases) have been detected. This protein probably also plays a role in sugar metabolism.

The intracellular glycosidases hydrolyze oligo- and disaccharides to monosaccharides such as glucose and xylose. Glucose appears to be degraded exclusively by the nonphosphorylating variant of the Entner-Doudoroff pathway (Schönheit and Schäfer 1995). At this stage, only a limited number of glycolytic enzymes has been identified in the *S. solfataricus* database: glyceraldehyde-3-phosphate dehydrogenase/phosphoglycerate kinase (GAPDH/PGK), non-phosphorylating glyceraldehyde-3-phosphate dehydrogenase (GAPN), and phosphoenolpyruvate synthase.

In *M. jannaschii, Methanobacterium thermoautotrophicum*, and *A. fulgidus* the presence of at least a gluconeogenesis pathway has been anticipated, and some archaea are probably able to degrade intracellular carbon reserves (e.g., glycogen in certain methanogens). Hence some glycolytic pathway is also to be expected. However, only a limited number of genes for the gluconeogenesis/glycolysis enzymes has been identified in the completed genomes (Bult et al. 1996; Klenk et al. 1997; Smith et al. 1997). Consequently, some of the enzymes involved in the central metabolic pathways may participate in minor or major deviations from versions of the pathways characterized in some Proteobacteria. It should be emphasized that variations on the classical bacterial theme have recently been described for archaeal glycolysis: two ADP-dependent hexose kinases (HK, PFK), a unique glyceraldehyde-3-phosphate converting enzyme (GAPOR) of *Pyrococcus furiosus* (Kengen et al. 1996; Van der Oost et al., in manuscript), and a PPi-dependent PFK and GAPN in *Thermoproteus tenax* (Siebers, Brunner, and Hensel, personal communication).

Repetitive elements in the *Sulfolobus solfataricus* genome

One of the distinctive features of the *S. solfataricus* genome is the large number of repetitive elements. There are several different kinds of these repetitive elements, and we briefly summarize their general features here.

The first group are the insertion elements (IS elements). Typically, these elements are specific to *S. solfataricus*; i.e., searches against the public databases do not reveal any other similar sequences in other species. To date, we have identified more than six IS elements that were previously unknown. All these IS elements contain ORFs that encode different proteins. All the known IS elements have terminal inverted repeats that are flanked by short direct repeats. Some of these elements are active in the growing *S. solfataricus* cultures, which have yielded cosmid clones differing in the location of an insertion element. The most prominent types of IS elements are ISC1217 (Schleper et al. 1994) and ISC1439 (Sensen et al. 1996). We expect the complete genome to have at least 15–20 copies of each of these two IS elements. The sequence similarity between the individual copies of the IS elements can vary (sometimes by more than 10% on the protein level) and some IS elements do not span the entire length of the "type" (i.e., the largest) copy. This suggests that these elements were inserted into the *S. solfataricus* genome over a long period of time and can decay after insertion. Some areas of the *S. solfataricus* genome contain high frequencies of various IS elements. In these regions, which can be between 50 and 100 kbp long, up to 95% of the sequence is repeated elsewhere in the genome. Similar transposases have been identified in the conjugative *Sulfolobus* plasmid pNOB8, suggesting that they may exchange between the genome and the plasmids (She et al., in press).

Gene families form a second type of repeated element. As in other sequenced genomes, certain gene families occur (e.g., ABC transporter genes), which form large clusters of similar, but not identical, gene sets. We expect the number of gene families to be similar to that in other archaeal genomes (Bult et al. 1996; Klenk et al. 1997; Smith et al. 1997).

Noncoding repeats are infrequent in the *S. solfataricus* genome, with the exception so far of two 6.5-kbp regions. The first region contains 48 copies of the sequence pattern GATTAATCCCAAAAGGAATTGAAAG, followed by 46 copies of GATTAATCC**T**AAAAGGAATTGAAAG (which varies from the first sequence only at the T in position 10). These repetitive elements are interrupted by nonrepetitive sequences of approximately 40 bp. About 14 kbp downstream, the second 6.5-kbp region shows 102 copies of the sequence pattern CTTTCAATTCCTTTTGGGATTAATC (there is no base exchange in this repeat), which is the reverse complementof the first kind of repeat. Again, the individual copies of the repeat are interrupted by approximately 40 bp of nonredundant sequence. Mojica et al. (1995) speculated that in *Haloferax mediterranei* and *Haloferax volcanii* these kind of repeats could be involved in replicon partitioning. Shorter arrays of similarly repeated sequences occur in the conjugative *Sulfolobus* plasmid pNOB8 (She et al., in press).

A third type of repetitive element considerably complicates sequence assembly. They are contained in genes that code for proteins that are constructed in the form of repetitive subunits. An example is a putative coiled-coil protein of *S. solfataricus* (Jeffries et al., in manuscript). The repeated sequences can be resolved only if their length is less than the maximum length of a sequence readout. Thus far, the longest gene detected that is organized in this way is less than 700 bp long and can be resolved on the LI-COR sequencers that produce an average readout length >850 bp.

Other genetic elements in *Sulfolobales*

In addition to the progress with the *S. solfataricus* genome, important parallel developments have been made in the characterization of extrachromosomal genetic elements from several Sulfolobales. The sequence of a virus, SSV1, is available (Reiter et al. 1987) and more recently a small helper satellite virus SSVX, has been sequenced (Zillig et al. this issue). Moreover, the small cryptic plasmids, pRN1 and pRN2, have been sequenced and analyzed (Keeling et al. 1996, in press) as well as a larger conjugative plasmid, pNOB8 (She et al., in press). In addition, many elements, both viral and plasmid, have been identified and remain to be sequenced. Details of these developments are summarized in an accompanying paper (Zillig et al., this issue). Suffice it to say here that a major effort is being made in several laboratories to exploit these viruses and plasmids for vector development using phenotypic marker genes. This includes genes for β-galactosidase (Elferink et al. 1996), selective marker genes for alcohol dehydrogenase (Aagaard et al. 1996; Aravalli and Garrett 1997) and a hygromycin-resistance enzyme (Cannio et al. 1996). These developments bode well for the future exploitation of the genome sequence for investigating the genetics, biochemistry, and physiology of *S. solfataricus*.

Concluding remarks

We are poised to complete the *S. solfataricus* genome-sequencing project within the next two years, with the bulk of sequencing completed in 1998. The project has already yielded many new insights into the organization and gene content of archaeal genomes, and it is expected that the completion of this genome will greatly enhance the understanding of phylogenetic relationships among the major phylogenetic domains.

Although we adopted a relatively laborious cosmid/λ cloning/sequencing strategy, the size and number of IS elements would make a total shotgun approach for *S. solfataricus* genome unlikely to succeed. Whenever two identical IS elements are identified in a contig, either this contig is subcloned to separate the identical regions, or

neighboring contigs that harbor only one instance of the repeat are used to guide the assembly of the affected contig. This approach seems to be successful and there is now a trend, especially for the smaller collaborative groups, to choose the combination of a shotgun and a primer-walking sequencing phase as a strategy for sequencing whole genomes.

Although the *S. solfataricus* genome will not be the first completed archaeal genome, it has led to many exciting and promising spinoffs that have extended the main project. These include the MAGPIE project (http://www.mcs.anl.gov/home/gaasterl/magpie.html) that was initiated using the *S. solfataricus* genomic data, the early stages of the Canadian Bioinformatics Resource (*http://www.cbr.nrc.ca/*), and interactions in the EU-supported Extremophile programs. These developments would not have taken place without the technical challenge of a whole genome sequencing project.

Acknowledgments The authors thank all the students and co-workers who have moved on to other tasks for their invaluable contributions to the success of the *Sulfolobus solfataricus* project. We would also like to thank Patrick Forterre (Université Paris Sud) and Willem de Vos (Wageningen Agricultural University) for their help in starting the European component of the *S. solfataricus* project. This project is supported by CGAT/Canada, CIAR/Canada, MRC/Canada, NRC/Canada and EU grant Bio 4CT960270. ACJ is sponsored through an NSERC visiting fellowship. This is NRCC publication No. 39778.

References

Aagaard C, Leviev I, Aravalli RN, Forterre P, Prieur D, Garrett RA (1996) General vectors for archaeal hyperthermophiles: strategies based on a mobile intron and a plasmid. FEMS Microbiol Revs 18:93–104

Aravalli RN, Garrett RA (1997) Shuttle vectors for hyperthermophilic archaea. Extremophiles 1:183–191

Baba T, Shinke R, Nanmori T (1994) Identification and characterization of clustered genes for thermostable xylan-degrading enzymes, β-xylosidase and xylanase, of *Bacillus stearothermophilus* 21. Appl Environ Microbiol 60:2252–2258

Bonfield JK, Smith KF, Staden R (1995) A new DNA sequence assembly program. Nucleic Acids Res 23:4992–4999

Bult CJ, White O, Olsen GJ, Zhou L, Fleischmann RD, Sutton GG, Blake JA, FitzGerald LM, Clayton RA, Gocayne JD, Kerlavage AR, Dougherty BA, Tomb JF, Adams MD, Reich CI, Overbeek R, Kirkness EF, Weinstock KG, Merrick JM, Glodek A, Scott JL, Geoghagen NSM, Weidman JF, Fuhrmann JL, Venter JC et al. (1996) Complete genome sequence of the methanogenic archaeon *Methanococcus jannaschii*. Science 273:1058

Cannio R, Contursi P, Rossi M, Bartolucci S (1996) An autonomously replicating, transforming vector for *Sulfolobus solfataricus*: expression in *Sulfolobus* cells of a mesophilic drug resistance adapted to the higher temperature. Abstracts, First international congress on extremophiles, Estoril, Portugal, p 35

Charlebois RL, Schalkwyk LC, Hofman JD, Doolittle WF (1991) Detailed physical map and set of overlapping clones covering the genome of the archaebacterium *Haloferax volcanii* DS2. J Mol Biol 222:509–524

Charlebois RL, Gaasterland T, Ragan MA, Doolittle WF, Sensen CW (1996) The *Sulfolobus solfataricus* P2 genome project. FEBS Lett 389:88–91

Charlebois RL, Sensen CW, Doolittle WF, Brown JR (1997) Evolutionary analysis of the *his* CGABdFDEH gene cluster from the archaeon *Sulfolobus solfataricus* P2. J Bacteriol 179:4429–4432

De Rosa M, Gambacorta A, Bullock JD (1975) Extremely thermophilic acidophilic bacteria convergent with *Sulfolobus acidocaldarius.* J Gen Microbiol 86:156–164

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

De Smet KLA, Jamil S, Stoker NG (1993) Tropist 3: a cosmid vector for simplified mapping of both G-C rich and A-T rich genomic DNA. Gene 136:215–219

Elferink MGL, Schleper C, Zillig W (1996) Transformation of the extremely thermoacidophilic archaeon *Sulfolobus solfataricus* via a self-spreading vector. FEMS Microbiol Lett 137:31–35

Gaasterland T, Sensen CW (1996a) MAGPIE: automated genome interpretation. Trends Genet 12:76–78

Gaasterland T, Sensen CW (1996b) Fully automated genome analysis that reflects user needs and preferences. A detailed introduction to the MAGPIE system architecture. Biochimie (Paris) 78:302–310

Gaasterland T, Lobo J, Maltsev N, Chen G (1994) Assigning function to CDS through qualified query answering. In: Altman R, Brutlag D, Karp P, Lathrop P, Searls D (eds) Proceedings of the second international conference on intelligent systems for molecular biology. AAAI Press, Cambridge, pp 348–353

Kaine BP (1987) Intron-containing genes of *S. solfataricus*. J Mol Evol 25:248–254

Kaine BP, Gupta R, Woese CR (1983) Putative introns in tRNA genes of prokaryotes. PNAS 80:3309–3312

Keeling PJ, Klenk H.-P, Singh RK, Schenk ME, Sensen CW, Zillig W, Doolittle WF (in press) *Sulfolobus islandicus* plasmids pRN1 and pRN2 share distant but common evolutionary ancestry. Extremophiles

Kengen SWM, Stams AJM, de Vos WM (1996) Sugar metabolism of hyperthermophiles. FEMS Microbiol Rev 18:119–137

Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, Dodson RJ, Gwinn M, Hickey EK, Peterson JD, Richardson DL, Kerlavage AR, Graham DE, Kyrpides NC, Fleischmann RD, Quackenbush J, Lee NH, Sutton GG, Gill S, Kirkness EF, Dougherty BA, McKenney K, Adams MD, Loftus B, Venter JC et al. (1997) The complete genome sequence of the thermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus.* Nature (Lond) 390:364–370

Maruta K, Mitsuzumi H, Nakada T, Kubota M, Chaen H, Fukuda S, Sugimoto T, Kurimoto M (1996) Cloning and sequencing of a cluster of genes encoding novel enzymes of trehalose biosynthesis from the thermophilic archaebacterium *Sulfolobus acidocaldarius*. Biochim Biophys Acta 1291:177–181

Mojica FJM, Ferrer C, Juez G, Rodriguez-Valera F (1995) Long stretches of short tandem repeats are present in the largest replicons of the Archaea *Haloferax mediterranei* and *Haloferax volcanii* and could be involved in replicon partitioning. Mol Microbiol 17:85–93

Prisco A, Moracci M, Rossi M, Ciaramella M (1995) A gene encoding a putative membrane protein homologous to the major facilitator superfamily of transporters maps upstream of the beta-glycosidase gene in the archaeon *Sulfolobus solfataricus*. J Bacteriol 177:1614–1620

Reiter W-D, Palm P, Henschen A, Lottspeich F, Zillig W, Grampp B (1987) Identification and characterization of the genes encoding three structural proteins of the *Sulfolobus* virus-like particle SSV1. Mol Gen Genet 206:144–153

Schleper C, Roder R, Singer T, Zillig W (1994) An insertion element of the extremely thermophilic archaeon *Sulfolobus solfataricus* transposes into the endogenous β-galactosidase gene. Mol Gen Genet 243:91–96

Schönheit P, Schäfer T (1995) Metabolism of hyperthermophiles. World J Microbiol Biotechnol 11:26–57

She Q, Phan H, Garrett RA, Albers S-V, Stedman, KM, Zillig W (in press) Genetic profile of pNOB8 from *Sulfolobus*: the first conjugative plasmid from an archaeon. Extremophiles

Sensen CW, Klenk HP, Singh RK, Allard G, Chan CC, Liu QY, Penny SL, Young F, Schenk ME, Gaasterland T, Doolittle WF, Ragan MA, Charlebois RL (1996) Organizational characteristics and information content of an archaeal genome: 156 kb of sequence from *Sulfolobus solfataricus* P2. Mol Microbiol 22:175–191

Smith DR, Doucette-Stamm LA, 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, Safer H, Reeve JN et al. (1997) Complete genome sequence of *Methanobacterium thermoautotrophicum* DH: functional analysis and comparative genomics. J Bacteriol 179:7135–7155