INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

Université d'Ottawa · University of Ottawa

Comparison of gene organization in the region that surrounds the 16S rRNA gene in seven different *Sulfolobales*

Anick De Moors

Thesis submitted to the
School of Graduate Studies and Research
University of Ottawa
in partial fulfillment of the requirements for the
M. Sc. degree in the
Ottawa-Carleton Institute of Biology

Ottawa, Ontario, Canada

July 2, 1998

In memory of my father

ABSTRACT

Stability in genome organization has been widely studied by comparing physical and genetic maps of different genomes and, by comparing complete sequences from different genomes. Many of these comparative studies between related species have shown some measure of conservation in gene organization. In order to determine if rearrangements in gene organization occur gradually over time, a set of related species needed to be compared in detail. My research consisted of a comparative sequence analysis of a selected region of the genome of various crenarchaeotal species belonging to the order *Sulfolobales*. More specifically, the gene organization of an equivalent 40-50 kbp region that surrounds the 16S rRNA gene of seven related strains or species was compared. The hypothesis of this research was that rearrangements in gene organization mirror phylogenetic history. Results from the analysis showed many blocks of conserved gene organization. A qualitative tree based on these similarities in gene order was constructed for comparison with the actual distance tree based upon the 16S rRNA sequences. My hypothesis was refuted on the basis that changes in gene organization did not mirror the phylogeny. Instead, my comparative analysis supported a punctuated model of genomic evolution. The use of gene order as a parameter to build phylogenies is discussed.

RÉSUMÉ

La stabilité dans l'organisation des génomes a été grandement étudiée en comparant les cartes physiques et génétiques de différents génomes et, en comparant les séquences complètes de différents génomes. Plusieurs de ces études comparatives entre espèces voisines, ont démontrées une certaine conservation dans l'organisation des gènes. Afin de déterminer si les réarrangements dans l'organisation des gènes surviennent graduellement en fonction du temps, plusieurs espèces voisines doivent être comparées. Ma recherche consiste en une analyse comparative des séquences appartenant à une région spécifique du génome de divers espèces crenarchaeotales faisant partie de l'ordre *Sulfolobales*. Plus précisément, l'organisation des gènes de la région de 40-50 kb qui entoure le gène ARN 16S ribosomale de sept espèces a été comparée. L'hypothèse de cette recherche énonce que les réarrangements dans l'organisation des gènes reflètent l'historique phylogénétique. Les résultats de l'analyse ont démontrés plusieurs blocs où l'organisation des gènes était conservée. Un arbre qualitatif basé sur les similarités dans l'ordre des gènes a été construis afin d'être comparé à l'arbre phylogénétique obtenu des distances calculées entre les séquences de l'ARN 16S ribosomale. Mon hypothèse a été refutée puisque les réarrangements dans l'organisation des gènes ne réfletent pas la phylogénie. Au lieu, mon analyse comparative soutient un modèle punctuel d'évolution génomique. L'utilisation de l'ordre des gènes comme paramètre pour construire des arbres phylogénétiques est discuté.

ACKNOWLEDGMENTS

I would like to thank my supervisor Dr. Robert Charlebois for giving me the opportunity to work in his laboratory. My master's research has been incredibly challenging. I have learned a great deal working under his supervision. I would like to further thank him for writing a program for the dot plots.

I would like to thank members of my research committee: Dr. George Carmody and Dr. Donal Hickey for their useful advices.

I would like to thank my coworkers Ghislaine Allard and Dave Sprott for their help and more importantly their support. They certainly created an enjoyable environment to work in. I'll always remember and cherish our numerous gossip sessions. I would also like to acknowledge other coworkers that contributed to the ambiance of the lab: Julie Chapados, Andrew St. Jean, Todd Monkman, Edith Lafleur, Neena Kushwaha and Christina Chan. I would like to extend special thanks to Andrea Bernhoff for her help during one summer.

I would like to thank my mother for her moral and financial support throughout my university years.

I would finally like to thank my fiancé Robert Fortier for his incredible patience and for his love.

TABLE OF CONTENTS

LIST OF FIGURES:

LIST OF TABLES:

INTRODUCTION

I. Gene organization

The chromosome contains the hereditary information necessary for cellular life. The typical primary structure of the procaryotic chromosome consists of a circular double-stranded DNA molecule tightly packed with contiguous genes. Gene organization is the arrangement of those genes along the chromosome. Gene organization is important since it may affect patterns of gene expression (Charlebois and St. Jean, 1995). In turn, the profile of gene expression will determine if a cell can adapt to an environment. Gene organization is the focus of this study. My research consists in comparing gene organization of seven related strains or species of *Sulfolobales* in a selected region of the genome. The null hypothesis of my research is that rearrangements in gene organization mirror phylogenetic history.

II. Stability of genomes

II.A. Rearrangements in gene organization

The genome evolves through three forms of genome changes: point mutations, genome rearrangements and lateral transfers. Point mutations operate at the nucleotide level whereas genome rearrangements and lateral transfers

operate on the structure of the chromosome. Thus, genome rearrangements and lateral transfers are responsible for altering gene organization.

The chromosome is capable of dynamic rearrangements. Recombination, transposition, deletion, insertion, duplication and inversion are different mechanisms that can change the spacing and in some cases, the order of genes in the genome (for reviews: Weinstock and Lupski, 1997; Hallet and Sherratt, 1997; Krawiec and Riley, 1990; Kolsto, 1997).

Recombination between pairs of homologous sequences is a major contributor to the plasticity of the genome. Insertion sequences, transposons, *rrn* loci and short sequence homologies are sites for recombination (Krawiec and Riley, 1990). Duplication and deletion occur when two repeated elements in the same orientation recombine, whereas inversion occurs when two repeated elements in the inverse orientation recombine. When a region of the genome is duplicated, more copies of genes are available. Thus, one copy of a gene can retain its original function, while the other copy can evolve and acquire a new function. When a section of the genome is deleted, the environment in that region of the genome is altered. Inversion is probably the most powerful force to change gene order. The effects of inversion on cell fitness have been studied (Hill *et al.*, 1990; Mahan *et al.*, 1990; Segall *et al.*, 1988). These studies have demonstrated that some intervals of the genome show inversion at high frequency (permissive)

while other intervals fail to show inversion (nonpermissive). Segall *et al.* (1988) have estimated that inversion occurred 100 times less frequently than duplication.

DNA sequences can be inserted into the genome without requiring any sequence homology. This type of rearrangement is important when considering gene organization because it can affect any region of the genome. Illegitimate recombination between sequences of little or no homology is an example of that type of rearrangement. Their mechanism although still unclear, involves breakage and reunion of DNA strands (Ehrlich *et al.*, 1993). Transposition of insertion sequences and transposons is also common in procaryotes and does not require sequence homology (for reviews: Chandler, 1997; Hallet and Sherratt, 1997). The status of a transposable element as a genetic locus is not as fixed compared to other genes in the genome. Transposing a genetic element from one location to another may have a dramatic effect or not on the cell.

Horizontal gene transfer represents another means to create diversity in the genome. In this event, genetic information from a genome is transferred to another. Horizontal transfer requires a vehicle to transport the genetic information and the machinery to insert the foreign DNA into the host genome. Lawrence and Ochman (1997) have estimated that 600 kb of the genome of *Escherichia coli* had been transferred horizontally.

The actual use of these rearrangements in the genome will depend on the species and on the environment where the organism is living. Changes in the

genome can be beneficial since it can promote rapid evolution of the organism in response to an environmental stress. Dybvig (1993) has reviewed instances where DNA rearrangements resulted in phenotypic switching needed for survival.

Rearrangements in the genome are also influenced by SOS repair and mismatch repair. SOS repair is expected to accelerate the rate of evolution since the mutation and recombination rates are increased. While on the contrary, mismatch repair is expected to reduce the rate of evolution since the mutation and recombination rates are decreased. An SOS response is important for bacterial survival in stressful situations (Matic *et al.*, 1995).

II.B. Forces that maintain gene organization

There are forces that are maintaining gene organization and preventing losing vital chromosomal identity. Some of these forces are: effective gene dosage, symmetry of the chromosome, local genetic context, direction of transcription, clustering of functionally related genes, and barrier to the formation of some rearrangements.

Gene dosage involves the position of genes relative to the origin of replication (*oriC*). Genes expected to be highly expressed might require to be located near the *oriC* to increase their copy numbers (Schmid and Roth, 1987). Gene dosage effect is the result of the presence of many replication forks moving from *oriC* to *terC* in the cell. Schmid and Roth (1987) have demonstrated that the location of genes affects their expression.

Symmetry of the chromosome is important to minimize the time required to replicate the genome. The origin of replication *oriC* and the termination of replication *terC* are thus separated by 180° to keep the two replicating arms the same length (François *et al.*, 1990; Hill *et al.*, 1990).

Gene expression is adapted to the local superhelical context (Charlebois and St. Jean, 1995). The chromosome in eubacteria is circular and divided into independent domains (Sinden and Pettijohn, 1981; Kellenberger, 1990). In each domain, negative supercoils are introduced by DNA gyrase (topoisomerase II) and relaxed by topoisomerase I (Drlica *et al.*, 1990). The global degree of supercoiling of a domain is determined by the expression of the topoisomerases (I and II), the presence of proteins that bind and wrap the DNA and, the presence of nicks during DNA replication and DNA repair. The orientation of genes also affects the global degree of supercoiling since the transcription bubbles are creating positive supercoils as bubbles progress. Thus, genes in the same orientation prevent jamming progress or denaturing the intergenic DNA. Promoters further have optimum degrees of supercoiling twist and writhe (Jordi *et al.*, 1995; Nickerson and Achberger, 1995; Jyothirmai, 1994). Some environmental stresses have shown to alter supercoiling, and consequently change patterns of gene expression (Higgins *et al.*, 1990; Dorman, 1996). Supercoiling changes have been observed in situations of change in osmolarity (Higgins *et al.*, 1988), change in oxygen tension (Dorman *et al.*, 1988), starvation (Balke and Gralla, 1987), entry into

stationary phase (Dorman *et al.*, 1988), and transition from one carbon source to another (Balke and Gralla, 1987). Altogether, genes seem to be adapted to their local level of supercoiling. Therefore, changing the order of genes implicates changing the context where genes work best (Charlebois and St. Jean, 1995).

The direction of transcription relative to the replication is also an important factor. Both DNA polymerase and RNA polymerase require a close contact with the DNA. When DNA polymerase moves in the opposite orientation of the RNA polymerase, head-on collision results in the interruption of the movement of the polymerases. This conflict can be prevented by avoiding them. Brewer (1990) and Blattner *et al.* (1997) have shown a strong bias in gene orientation in the *Escherichia coli* genome. The genes are transcribed 2.4 times more often in the same direction as DNA replication. The bias is even greater for genes that are transcribed actively. Kunst *et al.* (1997) have also shown a strong transcription orientation bias in the *Bacillus subtilis* genome. The genes are transcribed 75% of the time in the direction of replication.

The clustering of functionally related genes has been widely studied. A well known example is the cotranscriptional regulation that controls gene order in operons. For instance, the genes encoding L11, L1, L10 and L12 equivalent ribosomal proteins are clustered in the same order in archaea as in eubacteria (Shimmin *et al.*, 1989). Another instance, the organization of the spectinomycin operon of bacteria and archaea is similar (Scholzen and Arndt, 1991).

Finally, some rearrangements do not occur because of a barrier to their formation. For instance, rearrangements in the region of *terC* do not occur since it would disrupt DNA replication (Hill *et al.*, 1987; François *et al.*, 1990). Some other rearrangements are simply lethal to the organism because of the loss or inactivation of essential genes.

II.C. Genomic stability

An equilibrium must exist between the forces that are disrupting gene organization and the forces that are maintaining gene organization. The balance of these forces may also be different in various organisms. We can hypothesize that when the forces maintaining gene order are stronger than the forces disrupting gene order, genomic stability is achieved. Charlebois and St. Jean (1995) have suggested a model which predicts that genes are placed in genomic positions that optimize their expression in the different environments encountered by the cell. If a lineage was challenged by a new environment, the forces that maintain gene order would relax, genes could rearrange themselves, and a new stable genomic map would emerge. Each gene would further adapt to improve its performance.

Gene organization is without any doubt important for a genome. Nonetheless, it is also interesting to investigate gene organization between two different genomes. When two genomes have a similar gene organization, it implies that these organisms share a life history. While on the contrary, if two genomes have a dissimilar gene organization, it implies that at some point, the

environmental challenges encountered by these two organisms have differed. Genomic stability between two genomes has been widely studied in the last decade. Comparative mapping and comparative genomics are the two types of approaches that have been used to examine gene organization between genomes.

III. Comparative analyses of gene organization

III.A. Comparative genome mapping

A genome map is a graphic representation of the organization of the genome. Two types of maps can be produced: physical maps and genetic maps. They differ from each other by the way they are constructed. (for reviews: Fonstein and Hasselkorn, 1995; Fonstein and Hasselkorn, 1997; Cole and Saint Girons, 1994)

Physical maps can be built by a top-down or bottom-up approach. The top-down method involves separating large DNA fragments by pulsed-field gel electrophoresis (PFGE) using rare cutting enzymes. For instance, the intron endonuclease I-*Ceu*I which specifically cleaves the rrl (23S rRNA) genes in bacteria, is often used (McClellan *et al.*, 1997). The size and linkage of these fragments obtained are then determined. The bottom-up method involves the linkage of overlapping clones. This latter approach has a much finer resolution if one desires to characterize the genome of an organism. (Fonstein and Haselkorn, 1995) Regardless of the approach taken, most physical maps carry genetic

markers that are usually localized by hybridization. (Fonstein and Haselkorn, 1995; Cole and Saint Girons, 1994)

Genetic maps are not as used because they are labor-intensive and restricted to organisms for which genetic tools are available. The distances between genetic markers are measured by recombination frequencies. Consequently, they are sensitive to recombination hot spots. Genetic maps have only been constructed for *Escherichia coli* K-12 (Bachmann, 1983), *Salmonella typhimurium* LT2 (Sanderson and Roth, 1988), *Bacillus subtilis* (Piggot, 1990), *Streptomyces coelicolor* A3 (Hopwood and Kieser, 1990), *Staphylococcus aureus* (Pattee, 1990), *Pseudomonas putida* (Holloway *et al.*, 1990) and *Pseudomonas aeruginosa* (Holloway *et al.*, 1990).

Comparative mapping has been used to study genomic stability. The analysis involves comparing the order of genetic markers of two maps and determining the type and frequency of rearrangements since the divergence of the two organisms. Comparative mapping can be performed at three phylogenetic levels: intraspecies, interspecies or intergenic. Results from comparative mapping have shown various degrees of conservation in gene organization between strains, species or genera. Fonstein and Haselkorn (1995) have proposed two levels of conservation in gene organization in procaryotes; one level showing a strict conservation in gene order, and the second level showing no long range conservation of maps.

Conservation in gene order can be observed within the species *Escherichia coli*, *Salmonella typhimurium*, *Clostridium perfringens*, *Mycoplasma hominus*, *Lactococcus lactis*, *Streptomyces* spp., *Borrelia* spp., *Methanobacterium* spp. and *Haloferax* spp.. Conservation in gene order can also be observed between the species *Salmonella enteridis* and *Salmonella typhimurium* and; *Salmonella typhimurium* and *Escherichia coli*. Here is a summary of the conclusions drawn from these comparative analyses. Perkins *et al.* (1993) have compared maps from 6 strains of *Escherichia coli*, and identified 41 polymorphic loci which were mostly insertions and deletions of 1 to 86 kb. Liu and Sanderson (1995a) have studied the I-*Ceu*I skeleton of 17 strains of *Salmonella typhimurium* and observed a strong conservation in length and order. Canard *et al.* (1992) have studied 10 strains of *Clostridium perfringens*, and reported conservation in the order of genetic loci despite three hypervariable regions. Ladefoged and Christiansen (1992) have observed conservation in gene organization of five *Mycoplasma hominis* genomes, aside from one large inversion in one of the strains. Gene order of *Mycoplasma hominis* was also found to be identical to that of *Clostridium perfringens*. Le Bourgeois *et al.* (1995) have compared maps of *Lactococcus lactis* subsp. *cremoris* and *Lactococcus lactis* subsp. *lactis*, and noticed an overall conservation in gene organization with the exception of one large inversion of half of the genome. Leblond *et al.* (1993) have shown that the physical maps of *Streptomyces lividans* and *Streptomyces coelicolor* have a similar organization

except for one unstable region. Casjens *et al.* (1995) and Ojaimi *et al.* (1994) have demonstrated that the genetic organization of *Borrelia burgdorferi*, *Borrelia garinii*, *Borrelia afzelii*, *Borrelia japonica* and four other *Borrelia* isolates was identical. Stettler *et al.* (1995) have shown that the global arrangement of gene order of *Methanobacterium wolfei* is very similar to that of *Methanobacterium thermoautotrophicum*. López-García *et al.* (1995) studied the gene organization of *Haloferax volcanii* and *Haloferax mediterranei*, and concluded that the maps differed mainly by two inversions and a few translocations. Liu *et al.* (1993) have reported that gene order in *Salmonella enteritidis* and *Salmonella typhimurium* LT2 were identical except for one inversion. Riley and Sanderson (1990) have shown that the genetic maps of *Escherichia coli* K-12 and *Salmonella typhimurium* LT2 are similar, apart from one large inversion and a few deletions and insertions.

The group of microorganisms that are demonstrating highly rearranged chromosomes is composed of *Bacillus cereus*; *Helicobacter pylori*; *Rhodobacter capsulatus*; *Pseudomonas stutzeri*; *Leptospira interrogans*; *Salmonella typhi versus Salmonella typhimurium* and; *Haloferax volcanii* versus *Halobacterium salinarium*. Here are the main conclusions drawn from these comparative analyses. Carlson and Kolsto (1993) suggested that the taxonomy of the *Bacillus* genus needs reappraising since gene order of *Bacillus thuringiensis* and some strains of *Bacillus cereus* is similar, while gene order of two other strains of

Bacillus cereus differ as much as *Bacillus subtilis* does. Jiang *et al.* (1996) have reported no conservation in gene organization and no clustering of genes within five *Helicobacter pylori* strains. Nikolskaya *et al.* (1995) have studied a 1.2 Mb chromosomal region of three *Rhodobacter capsulatus* strains that showed numerous large and small rearrangements. Ginard *et al.* (1997) detected diversity in chromosomal arrangements in 20 strains of *Pseudomonas stutzeri*. Zuerner *et al.* (1993) compared maps of the larger replicon of two strains of *Leptospira interrogans* and demonstrated many large rearrangements. Liu and Sanderson (1995b and 1996), have shown that the genome of *Salmonella typhi* Ty2 doesn't share the same organization of *Salmonella typhimurium* and *Escherichia coli*. For this matter, *Salmonella typhi* Ty2 has undergone major genome rearrangements such as homologous recombination between *rrn* genes. Finally, St. Jean and Charlebois (1996) reported different chromosomal arrangements in *Haloferax volcanii* DS2 and *Halobacterium salinarium* GRB.

Comparative genome mapping has been useful in studying genomic stability mainly because the analysis is simple, rapid and inexpensive. Comparative mapping has also been used for typing of strains (Lück *et al.*, 1995). The main weakness of comparing maps is the lack of resolution. The resolution of comparative mapping depends on the approach taken to construct the maps and the number of markers that are actually being compared between the two organisms. The analysis can determine to a certain level, the degree of conservation in gene

organization, but doesn't identify the forces involved. Furthermore, genomic stability versus instability is subjective and poorly quantified. Comparative mapping is an important study, but remains one of the preliminary stages to understanding stability in gene organization.

II.B. Comparative genomics

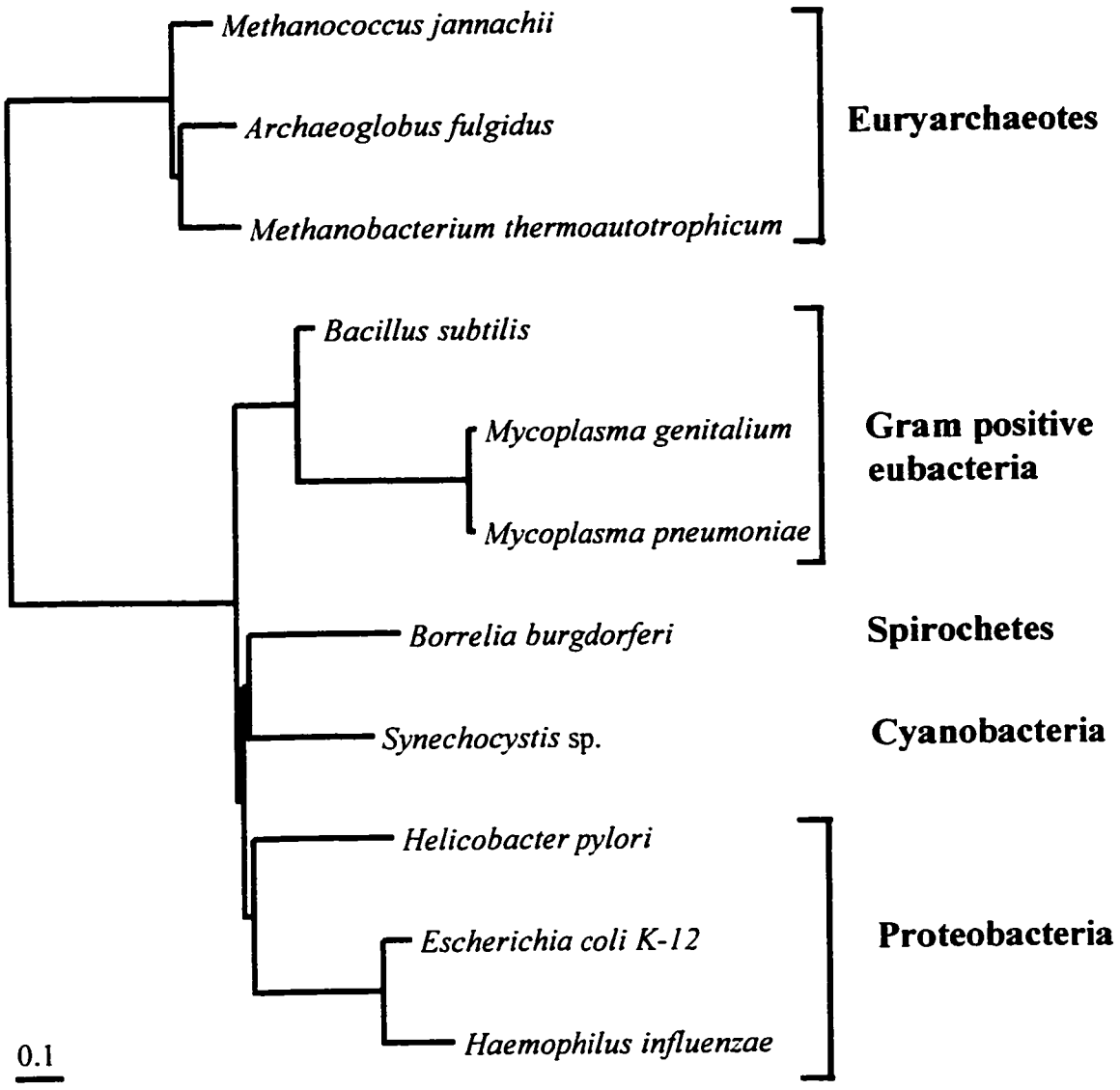
Comparative genomics has also been used to measure map stability. In this type of analysis, two genomes that have been completely sequenced are compared to each other. Whole procaryotic genomes have been sequenced for *Haemophilus influenzae* (Fleischmann *et al.*, 1995), *Mycoplasma genitalium* (Fraser *et al.*, 1995), *Synechocystis* sp. (Kaneko *et al.*, 1996), *Methanococcus jannaschii* (Bult *et al.*, 1996), *Mycoplasma pneumoniae* (Himmelreich *et al.*, 1996), *Methanobacterium thermoautotrophicum* (Smith *et al.*, 1997), *Helicobacter pylori* (Tomb *et al.*, 1997), *Bacillus subtilis* (Kunst *et al.*, 1997), *Archaeoglobus fulgidus* (Klenk *et al.*, 1997), *Borrelia burgdorferi* (Fraser *et al.*, 1997), *Escherichia coli* (Blattner *et al.*, 1997) and *Aquifex aeolicus* (Deckert *et al.*, 1998). Moreover, approximately 50 other procaryotic genomes are in the process of being sequenced.

Comparative genomics has the advantage of having a high resolution over comparative mapping. However, the genomes that have been completely sequenced so far, are often too distantly related to observe any conservation in gene order. Clearly, sequencing other genomes that are more closely related

would be more informative, but also more labor intensive, time-consuming and expensive. The phylogenetic relationship between the different genomes that are available is shown in figure 1.

Most comparative genomic analyses performed until now, have demonstrated little conservation in gene organization within genomes (for review: Kolsto, 1997). Tatusov *et al.* (1996) have compared the genomes of *Haemophilus influenzae* and *Escherichia coli*, two γ-Proteobacteria. No large-scale conservation of gene organization was observed between the two genomes. Nevertheless, many short conserved gene strings were discovered. Fewer than 50% of these strings are known as operons. Tamames *et al.* (1997) have also compared the genomes of *Haemophilus influenzae* and *Escherichia coli* by performing statistics on neighboring pairs of genes. They have shown a strong tendency for genes within the same functional class to cluster together. From a total of 800 neighbors in *Haemophilus influenzae* and 1971 in *Escherichia coli*, 291 stayed neighbors in both species. Furthermore, the relative direction of transcription was found to be conserved between homologous gene pairs (Tamames *et al.*, 1997). Fraser *et al.* (1995) have compared the genomes of *Mycoplasma genitalium*, a Gram Positive bacterium and *Haemophilus influenzae*, a γ-Proteobacterium. No long-range conservation in gene order was observed. However, short regions of conserved gene order were observed, particularly two clusters of ribosomal proteins. Mushegian and Koonin (1996) have further compared *Escherichia coli*,

Figure 1. Phylogenetic tree relating the different procaryotes whose genomes have been completely sequenced. The tree was obtained from The Ribosomal Database Project site (http://rdpwww.life.uiuc.edu/), based upon the small-subunit rRNA sequences under maximum likelihood's model. Bar, 0.1 substitutions/site.


Methanococcus jannachii
Archaeoglobus fulgidus
Methanobacterium thermoautotrophicum
Euryarchaeotes
Bacillus subtilis
Mycoplasma genitalium
Mycoplasma pneumoniae
Gram positive eubacteria
Borrelia burgdorferi
Spirochetes
Synechocystis sp.
Cyanobacteria
Helicobacter pylori
Escherichia coli K-12
Haemophilus influenzae
Proteobacteria
0.1

Mycoplasma genitalium and *Haemophilus influenzae* to each other. Only 27 gene strings were found to be conserved between the three bacterial species, and 23 of these are known to be operons. Siefert *et al.* (1997) have compared gene order of four bacteria, *Haemophilus influenzae*, *Mycoplasma genitalium*, *Mycoplasma pneumoniae* and *Synechocystis* spp., and one archaeon, *Methanococcus jannachii*. Little conservation in gene organization was observed between the different genomes. Only 16 gene strings were conserved among the four bacteria and only 8 of these are conserved in archaea as well. Smith *et al.* (1997) have investigated gene organization in *Methanococcus jannachii* and *Methanobacterium thermoautotrophicum*, two euryarchaeotes. Only 14% of orthologous genes had the same neighbor in the two genomes.

Conservation in gene organization can however be observed when the genomes of closely related organisms are compared. Himmelreich *et al.* (1997) compared the genomes of two *Mycoplasma* species: *Mycoplasma genitalium* and *Mycoplasma pneumoniae*. The two genomes were subdivided into 6 segments. Within these segments, gene order was well conserved. However, the order of these 6 segments was different between the two genomes. These rearrangements can be explained by homologous recombination between the segments. The translocations did not however interfere with the orientation of transcription relative to replication.

III.C. Comparative sequencing analysis of a selected region of many genomes

The null hypothesis of my research is that rearrangements in gene organization mirror phylogenetic history. In order to test my hypothesis, changes in gene organization relative to sequence divergence were compared in seven closely related organisms. Because of the lack of resolution of comparative mapping, a high resolution comparison at the sequence level was required. Furthermore, in order to make this analysis realistic and feasible, my comparison was limited to a specific region of the genome. More specifically, the 40-50 kbp region surrounding the 16S rRNA locus of seven related species and strains of *Sulfolobales* was compared. The 16S rRNA gene was chosen as a reference point in the genome because of its high conservation and because a single copy is present in the *Sulfolobales*. The *Sulfolobales* were chosen for my comparison since my laboratory was already involved in the genome sequencing project of *Sulfolobus solfataricus* P2 (Sensen *et al.*, 1997). The sequencing of a genome can generate an enormous amount of data that can be very useful to study gene order in-depth. Therefore, I will be using some of the data generated for my gene organization study by using *Sulfolobus solfataricus* P2 as a reference strain. This type of comparative sequencing analysis in which a specific region of the genome is compared between many related organisms at the sequence level, represents the first of its kind.

IV. The *Sulfolobales*, sulfur-dependent thermoacidophilic archaea

All living organisms are divided into three domains based on the phylogeny of the rRNA: *Archaea*, *Bacteria* and *Eucarya* (Woese and Fox, 1977; Woese *et al.*, 1990). Figure 2 is a phylogenetic tree of life based on the small-subunit rRNA sequences, showing the three domains (Pace, 1996). The archaea and eubacteria are considered procaryotes primarily because they lack a nuclear membrane. The archaea share some features with the bacteria in the overall genome design, such as the presence of a single circular chromosome and gene organization. However, they differ from one another in a number of ways. (for reviews: Keeling *et al.*, 1994; Keeling and Doolittle, 1995)

The archaeal domain is divided into two main divisions on the basis of the small-subunit rRNA sequence: the *Euryarchaeota* and the *Crenarchaeota*. The euryarchaeotes consist of methanogens, extreme halophiles, sulfate-reducing species (the genus *Archaeoglobus*) and two types of thermophiles (the genus *Thermoplasma* and the Thermococcales). The crenarchaeotes consist of sulfur-dependent extreme thermophiles. All thermoacidophiles in the crenarchaeote division belong to the order of *Sulfolobales*. Figure 3 shows a phylogenetic tree based on the 16S rRNA gene showing the *Sulfolobales*. (Fuchs *et al.*, 1996) Within the *Sulfolobales*, the genera *Sulfolobus*, *Acidianus*, *Stygiolobus*, *Desulfurolobus* and *Metallosphaera* have been described. They mainly inhabit

Figure 2. Universal phylogenetic tree showing the three domains; the *Archaea*, the *Bacteria* and the *Eucarya*. The tree was reproduced from Pace (1996) and was based upon the small-subunit rRNA gene.

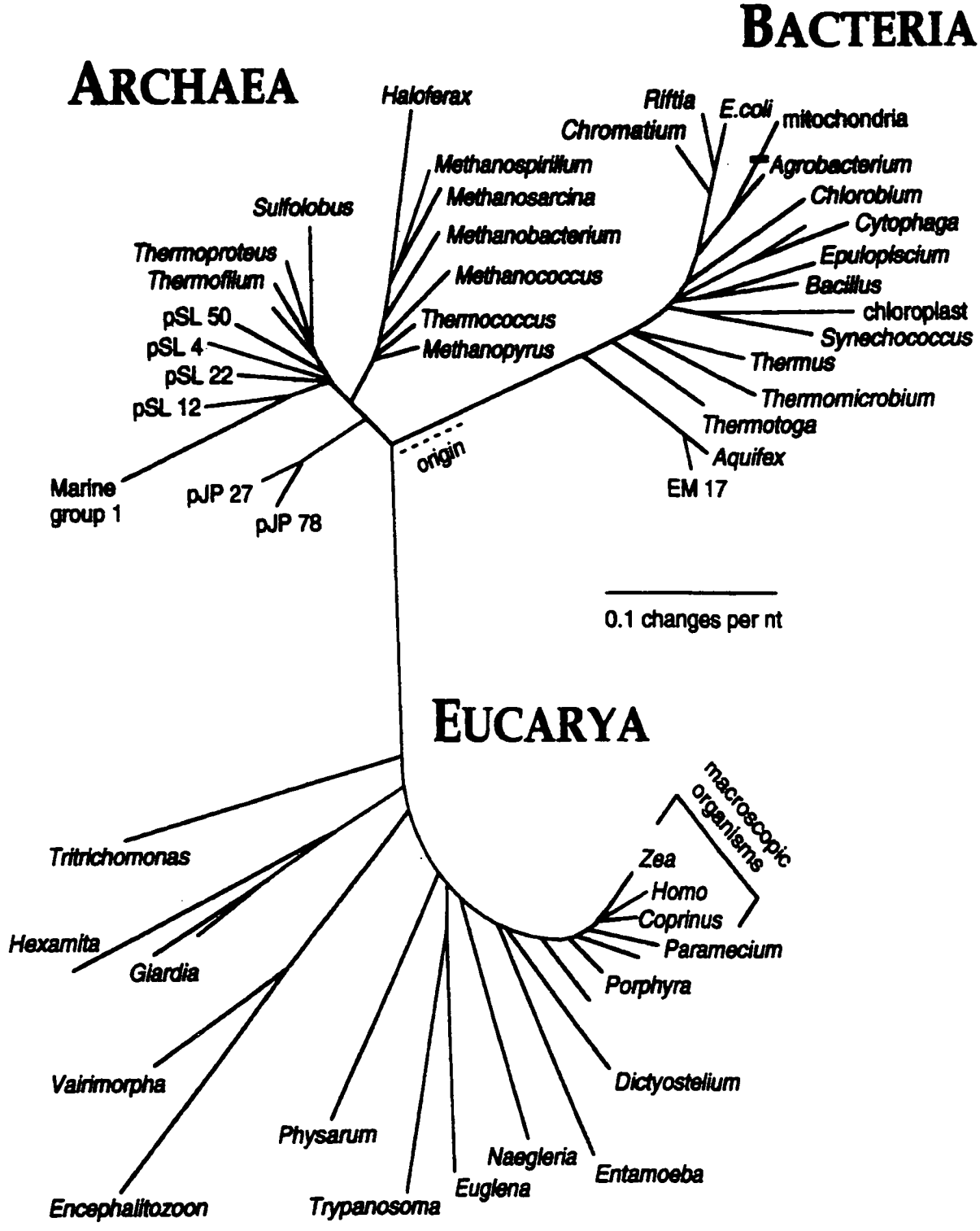
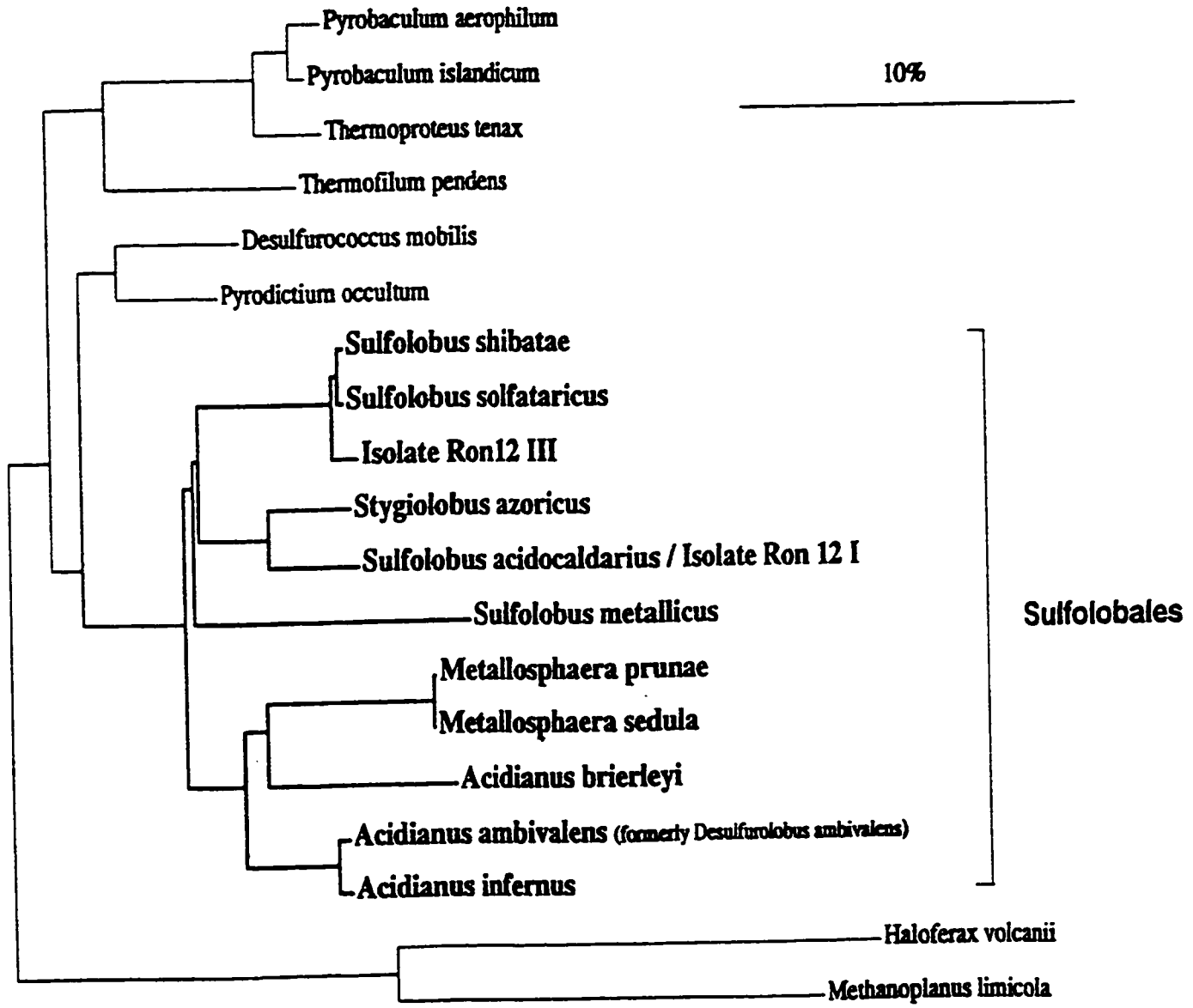

BACTERIA
ARCHAEA
Haloferax
Riftia
E.coli
mitochondria
Chromatium
Methanospirillum
Agrobacterium
Methanosarcina
Chlorobium
Sulfolobus
Cytophaga
Methanobacterium
Thermoproteus
Epulopiscium
Thermofilum
Methanococcus
Bacillus
pSL 50
Thermococcus
chloroplast
pSL 4
Methanopyrus
Synechococcus
pSL 22
Thermus
pSL 12
Thermomicrobium
Thermotoga
origin
Aquifex
Marine
group 1
pJP 27
EM 17
pJP 78
0.1 changes per nt
EUCARYA
macroscopic
organisms
Tritrichomonas
Zea
Homo
Hexamita
Coprinus
Giardia
Paramecium
Porphyra
Vairimorpha
Dictyostelium
Physarum
Naegleria
Entamoeba
Euglena
Encephalitozoon
Trypanosoma

Figure 3. Phylogenetic tree showing the *Sulfolobales*, based upon the 16S rDNA sequences of 12 thermoacidophilic members of archaea. The tree was reproduced from Fuchs *et al.* (1996). The distances have been calculated assuming Jukes and Cantor's model of evolution and the dendrogram was constructed by the least squares distance method of De Soete. Bar, 10 substitutions per hundred nucleotides.


Pyrobaculum aerophilum
Pyrobaculum islandicum
10%
Thermoproteus tenax
Thermofilum pendens
Desulfurococcus mobilis
Pyrodictium occultum
Sulfolobus shibatae
Sulfolobus solfataricus
Isolate Ron12 III
Stygiolobus azoricus
Sulfolobus acidocaldarius / Isolate Ron 12 I
Sulfolobus metallicus
Sulfolobales
Metallosphaera prunae
Metallosphaera sedula
Acidianus brierleyi
Acidianus ambivalens (formerly Desulfurolobus ambivalens)
Acidianus infernus
Haloferax volcanii
Methanoplanus limicola

continental sulfur-rich springs, share the ability to oxidize elemental sulfur and grow at low pH and high temperature. The different genera within that group can be distinguished by their metabolic and biochemical properties, their G+C content (31-45%), DNA/DNA hybridization and their 16S rRNA gene sequence (Grogan, 1989; Grogan *et al.*, 1990; Huber *et al.*, 1989). *Sulfolobus*, *Acidianus* and *Metallosphaera* are aerobes whereas *Desulfurolobus* and *Stygiolobus* are strict anaerobes. The strains chosen for my study were *Sulfolobus solfataricus* P2, *Sulfolobus solfataricus* P1, *Sulfolobus solfataricus* MT4, *Sulfolobus shibatae*, *Sulfolobus acidocaldarius*, *Sulfolobus tokodaii* and *Metallosphaera sedula*. These different organisms were estimated to have diverged from each other at different times, varying from 25 to 550 million years, based on the 16S rRNA sequence and a clock rate of 1% substitutions per 50 million years (Ochman and Wilson, 1987). The geographical origins of these organisms and their G+C contents are indicated in Table 1. In addition to being able to oxidize sulfur for energy, these *Sulfolobales* are facultative autotrophs and therefore, can also grow on rich media. The size of the genomes of *S. solfataricus* P2 and *S. tokodaii* have been estimated to be approximately 3 Mbp and 2.8 Mbp respectively (Sensen *et al.*, 1996: Kondo *et al.*, 1993). Many insertion sequences have been identified in *S. solfataricus* P2 (Sensen *et al.*, 1996). When hybridizing genomic DNA from the other *Sulfolobales* studied, to the insertion sequences found in *S. solfataricus* P2, most were also found to be present in *S. solfataricus* P1, *S. solfataricus* MT4 and

S. shibatae. However, very few were found in *S. acidocaldarius*. *S. tokodaii* and *M. sedula* at the level of stringency used (Unpublished data).

Table 1. Geographical origins of the *Sulfolobales* and their G+C contents.

	Geographical origin	G+C content
***S. solfataricus* P1**	Volcanic hot springs, Italy, Campi flegrei (Zillig *et al.*, 1980)	35% (Zillig *et al.*, 1980)
***S. solfataricus* P2**	Volcanic hot springs, Italy, Campi flegrei (Zillig *et al.*, 1980)	35% (Zillig *et al.*, 1980)
***S. solfataricus* MT4**	Hot water ponds, Italy, Naples, Agnano, Pisciarelli Solfatara (De Rosa *et al.*, 1975)	35% (Zillig *et al.*, 1980)
S. shibatae	Geothermal mud hole, Japan, Kiushuisland (Grogan *et al.*, 1990)	35% (Grogan *et al.*, 1990)
S. acidocaldarius	Acid hot spring, USA, Yellowstone National Park (Brock *et al.*, 1972)	38% (Zillig *et al.*, 1980)
S. tokodaii	Acidic spa, Beppu hot springs, Japan (Wakagi and Oshima, 1985)	40% (Yamagashi and Oshima, 1990)
M. sedula	Hot water ponds, Italy, Naples, Agnano, Pisciarelli Solfatara (Huber *et al.*, 1989)	45% (Huber *et al.*, 1989)

MATERIALS AND METHODS[1]

I. Growth of *Sulfolobales*

I.A. Strains

Sulfolobus solfataricus P2 (DSM 1617), *Sulfolobus solfataricus* P1 (DSM 1616), *Sulfolobus solfataricus* MT4 (DSM 5833), *Sulfolobus shibatae* (DSM 5389), *Sulfolobus acidocaldarius* (DSM 639) and *Metallosphaera sedula* (DSM 5348) were obtained from Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSM, Braunschweig, Germany). *Sulfolobus tokodaii* (formerly known as *Sulfolobus acidocaldarius* strain 7) was obtained from Akihiko Yamagishi (Tokyo Institute of Technology, Yokohama).

I.B. Culture conditions

The various strains were cultivated in Allen's (1959) medium, modified by Brock *et al.* (1972). This basal salts medium contains (per liter): 1.3 g $(NH_4)_2SO_4$, 280 mg KH_2PO_4, 250 mg $MgSO_4 \cdot 7H_2O$, 70 mg $CaCl_2 \cdot 2H_2O$, 20 mg $FeCl_3 \cdot 6H_2O$, 4.5 mg $Na_2B_4O_7 \cdot 10H_2O$, 1.8 mg $MnCl_2 \cdot 4H_2O$, 0.22 mg $ZnSO_4 \cdot 7H_2O$, 0.05 mg $CuCl_2 \cdot 2H_2O$, 0.03 mg $VOSO_4 \cdot 2H_2O$, 0.03 mg $Na_2MoO_4 \cdot 2H_2O$ and 0.01 mg $CoSO_4$. The basal salts medium was supplemented with 0.1% (w/v) yeast extract and 0.2% (w/v) sucrose unless stated otherwise. The pH was adjusted with H_2SO_4.

[1] Refer to Appendix I for recipes of solutions and media.

Solid media were prepared by adding (per liter) 8 g Gelrite gellan gum, 3 mg $CaCl_2$ and 10 mg $MgCl_2$ to this yeast extract-sucrose medium. The pH was adjusted to 3.5 with H_2SO_4. *S. solfataricus* P1, *S. solfataricus* P2, *S. shibatae* and *S. tokodaii* were grown on plates at 70°C, for 7 days. A single colony from these organisms was used to inoculate liquid medium. *S. acidocaldarius*, *S. solfataricus* MT4 and *M. sedula* were grown in liquid cultures only.

Liquid cultures were grown in glass hybridization tubes using 50 ml of medium. The cultures were incubated aerobically for 3-4 days in a hybridization oven. The different organisms were grown in their optimal growth conditions. *M. sedula* was grown at 65°C; *S. acidocaldarius* at 70°C; *S. solfataricus* P1, *S. solfataricus* P2, *S. tokodaii* and *S. shibatae* at 78°C; and *S. solfataricus* MT4 at 85°C. The pH of the cultures was adjusted to 2.5 for *M. sedula*; 3.0 for *S. acidocaldarius* and *S. tokodaii*; and 3.5 for *S. shibatae*, *S. solfataricus* P1, *S. solfataricus* P2 and *S. solfataricus* MT4. *S. solfataricus* P1 was supplemented with 0.1% yeast extract and 0.2% maltose whereas *M. sedula* was supplemented with 0.1% yeast extract only.

I.C. Storage

The cultures were stored at -80°C based on a protocol provided by Deutsche Sammlung von Mikroorganismen und Zellkulturen (DSM, Braunschweig, Germany). Briefly, 10 ml of culture was kept at room temperature for 1 hour then $CaCO_3$ (~200 mg) was added in excess. The culture was then left

at room temperature for 30 minutes to allow the $CaCO_3$ and $CaSO_4$ to settle to the bottom. The supernatant was transferred to another tube and centrifuged for 7 minutes. The cells were resuspended in 1 ml of fresh medium that had been neutralized with $CaCO_3$ and which contained 10% glycerol. The cells were frozen at -80°C.

II. Extraction of genomic DNA

Total genomic DNA from the various strains, was isolated from a 40 ml culture in the early to mid exponential phase of growth. The cells were centrifuged for 7 minutes. Cell pellets were resuspended in 200 µl of TE (10 mM Tris·Cl pH 8 and 1 mM EDTA). The cells were lysed by adding 20 ml of STE (1% Lauryl sarcosine, 50 mM Tris pH 8 and 50 mM EDTA) followed by gentle swirling. To allow complete lysis, the mixture was left at room temperature for 30 minutes. Three phenol : chloroform (1:1) extractions were performed, followed by an ethanol precipitation (1/10 volume of 3 M sodium acetate pH 5.5 and 2 volumes of 95% ethanol). The DNA was washed in 70% ethanol, air dried partially and suspended in 400 µl of TE.

III. Construction of cosmid libraries

III.A. Preparation of vector

Tropist3 was the cosmid vector used to construct the cosmid libraries (De Smet *et al.*, 1993). Twenty-five µg of Tropist3 was cut with *Sca* I (30 U) for 3

hours at 37°C between the two *cos* sites. The linearized vector was phenol : chloroform (1:1) extracted, ethanol precipitated, washed and resuspended in 50 µl of TE. A sample (1 µg) was run on a 0.8% TAE (50 mM Tris, 2 mM EDTA and 20 mM sodium acetate) agarose gel to confirm complete digestion of Tropist3. The 5' phosphorylated ends of the linearized vector were dephosphorylated by treatment with alkaline phosphatase (1 U) at 37°C for 30 minutes. The enzyme was denatured by incubating the vector at 65°C for 15 minutes. The vector was then phenol : chloroform (1:1) extracted, ethanol precipitated, washed, and resuspended in 87 µl of TE. Tropist3 was then cut with the cloning enzyme (30 U of *Hind* III, *Xba* I or *Bam* HI) for 3 hours at 37°C. Again, the vector was phenol : chloroform (1:1) extracted, ethanol precipitated, washed, and resuspended in 50 µl of TE. The final concentration of the prepared vector was 0.5 µg/µl. One µg of cut vector was self-ligated using 0.3 U of DNA T4 ligase at 12°C for 16 hours. Ligated and non-ligated vectors were verified on a 0.8% TAE agarose gel.

III.B. Preparation of partially digested genomic DNA

An ideal cosmid library contains all the genomic sequences at the same frequency. An important factor that influences this frequency is the choice of cloning enzyme. This restriction endonuclease should cut the genome randomly. *Hind* III, *Xba* I and *Bgl* II are suitable cloning enzymes since they frequently cut the genome of *Sulfolobales*. To determine which enzyme to use in the preparation of the cosmid library, 2 µg of genomic DNA was digested with the three

restriction enzymes and verified on a 0.6% TAE agarose gel. The cloning enzyme that cut the genome more frequently was chosen. The size of the biggest fragment should not be greater than 15 kb. *Xba* I was the enzyme chosen for *S. solfataricus* MT4 and, *Hind* III was chosen for *S. solfataricus* P1, *S. shibatae*, *S. acidocaldarius*, *S. tokodaii* and *M. sedula*.

Ten μg of genomic DNA was partially digested for 30 minutes at 37°C with different amount of cloning enzyme (8U, 6U, 4U, 3U, 2U, 1.5U, 1U and 0.5U). Samples of the partial digests (1 μg), a complete digest, uncut genomic DNA and uncut lambda DNA were loaded onto a 0.6% TAE agarose gel. The partials that had their main size range between 40 and 60 kb were chosen for preparing the cosmid library. The partials were treated with alkaline phosphatase (1U) at 37°C for 30 minutes to avoid ligation of non-contiguous DNA fragments. The enzyme was deactivated by adding EDTA to a final concentration of 20 mM. The partials were phenol : chloroform (1:1) extracted, ethanol precipitated, washed and resuspended in 16 μl of TE.

III.C. Ligation

The phosphatased partials (6 μg) were ligated to the prepared Tropist3 (0.5 μg) using one unit of T4 DNA ligase at 12°C for 18 hours. The ligation mix was stored in the freezer.

III.D. In vitro Packaging

The introduction of cosmids into bacterial cells was done by packaging the cosmids into lambda particles followed by the infection of host cells. The *in vitro* packaging was performed using Gigapack II XL packaging extract from Stratagene. The procedure was explained in the instruction manual. Briefly, 4 µl of ligation mix was added to a Freeze/Thaw extract tube, followed by 15 µl of Sonic extract. The packaging of cosmids was done at room temperature for 2 hours. Five hundred µl of phage dilution buffer was added, followed by 20 µl of chloroform.

III.E. Infection of host cells

E. coli ED8767 was the strain used as the host. *E. coli* was grown on a YT plate at 37°C for 18 hours. Ten ml of YT broth was inoculated with a single colony and incubated at 37°C for 18 hours. Fifty ml of YT broth supplemented with 0.2% maltose and 10 mM $MgSO_4$, was inoculated with 0.5 ml of *E. coli* culture and grown at 37°C with shaking for 4-5 hours ($O.D._{600}$ between 0.6 and 1). The cells were centrifuged for 10 minutes and resuspended in 10 mM $MgSO_4$ to an OD_{600} of 0.5.

One hundred fifty µl of packaged cosmids and 1.5 ml of prepared host cells were incubated for 20 minutes at 37°C with gentle shaking. The cells were grown at 37°C for 40 minutes with gentle shaking, by adding 3.75 ml of prewarmed YT broth. The cells were centrifuged and resuspended in 1.3 ml of YT broth. One

hundred µl aliquots were spread onto YT-kan plates and incubated at 37°C for 23 hours. Only the cells that were infected with a cosmid grew on such plates.

III.F. Storage of cosmid libraries

Colonies (800 to 1200) were picked using sterile toothpicks and grown in 100 µl YT-kan broth in 96-well lidded microtiter dishes at 37°C for 20 hours. The cultures were frozen at -80°C with 15% v/v glycerol.

IV. Screening of the cosmid libraries

IV.A. Amplification of the 16S rRNA gene to use as a probe

Based on a 16S rRNA gene alignment of various *Sulfolobales*, a set of primers was designed to amplify a part of the 16S rRNA gene. The chosen primers were 5'—TGTCAGCCGCCGCGGTAATACC—3' and 5'—GGAGTCGGAGGGTACGATTACG—3'. The PCR reaction was performed using 10 and 100 ng of genomic DNA, 4 mM $MgCl_2$, 200 µM dNTP, 30 pmol of each primer, 1X Thermo buffer and 1 U of Taq polymerase. The thermal cycling using a Gene Amp PCR system 9600 was set as follows: 95°C for 5 minutes; 30 cycles of: 1 minute at 94°C, 2 minutes at 65°C and 5 minutes at 72°C; and finally hold at 4°C. One third of the PCR reaction was run on a 1% TAE agarose gel.

IV.B. Preparation of blots

Before freezing the microtiter dishes at -80°C, a 96-pronged device was used to stamp the cultures onto GeneScreen Plus membranes (Dupont / NEN)

overlaid on top of YT-kan agar in large square dishes. Each microtiter dish was stamped twice to obtain duplicates of the cosmid libraries. The plates were incubated at 37°C for 20 hours. The blots were sequentially treated 5 minutes to the 4 following solutions: 1) 10% SDS lysis solution, 2) denaturing solution (0.5M NaOH and 1.5M NaCl), 3) neutralizing solution (0.5M Tris-Cl pH 7.4 and 1.5M NaCl), and 4) 2X SSC solution (3M NaCl and 0.3M sodium citrate). The colony blots were air dried and the DNA was fixed onto the membrane by autoclaving the blots for 1 minute. Just prior to the hybridization, the blots were soaked in a prewash solution (0.5% SDS, 1mM EDTA and 5X SSC) for 1 hour at 50°C. Colony debris was scrubbed off the blots using Kimwipes.

IV.C. Hybridization of the cosmid libraries

The colony blots were placed in a hybridization tube. The blots were prehybridized at 38°C for at least 1 hour with 15 ml of hybridization solution (1M NaCl, 50 mM Tris·Cl pH 7.6, 5% SDS and 50% v/v formamide) and 300 µl of 3 mg/ml denatured and sheared herring sperm DNA.

One µl of random hexamer primers (2.5µg/µl) was added to 300 ng of PCR-amplified 16S rRNA fragments, bringing the final volume to 17 µl with water. The DNA fragments and hexamers were boiled 5 minutes and quick chilled on ice. Three µl of BSA (1 mg/ml), 2.5 µl of 10X RP-C buffer, 2 µl of [α-^{32}P]dCTP (20 µCi) and 0.5 µl (1 U) of Klenow enzyme were added. The random-priming labeling was performed at 37°C for 45 minutes. The probe was then

precipitated with two volumes of 0.5 M ammonium acetate in 95% ethanol, centrifuged for 10 minutes and resuspended in 100 µl of water. The probe was boiled for 5 minutes, quick chilled and added to the hybridization tube containing the colony blots. The blots were hybridized overnight at 38°C.

The hybridization solution was drained and the blots were rinsed twice with 2X SSC / 1% SDS. The hybridization tube was half filled with prewarmed 2X SSC / 1% SDS and incubated at 65°C for 1 hour. The wash solution was drained and the blots were rinsed again with 2X SSC. The blots were placed in an autoradiography cassette between sheets of plastic wrap and exposed to a Kodak X-OMAT film with an intensifying screen at -80°C.

V. Choosing positive cosmid for preparation of plasmid library

V.A. Alkaline extraction of cosmids

The colonies that were positive on the autoradiogram, were grown on YT-kan plates for 22 hours at 37°C. The smallest colonies on plates were used to inoculate 15 ml of YT-kan broth. The cultures were grown with shaking at 37°C for 21 hours. The cells were spun 7 minutes and resuspended in 200 µl of TEG (20 mM Tris·Cl pH 8, 50 mM EDTA and 1% glucose). The cells were lysed when adding 400 µl of 0.2 M NaOH / 1% SDS and mixing by rocking. Three hundred µl of 7.5 M ammonium acetate kept at -20°C was added. The neutralized lysate was mixed and centrifuged for 7 minutes. The supernatant was extracted with

phenol : chloroform (1:1), precipitated with 500 µl of isopropanol, and resuspended in 300 µl of TE. The DNA was precipitated once more by adding 2 volumes of 0.5 M ammonium acetate in 95% ethanol, washed with 70% ethanol and resuspended in 50 µl of water.

V.B. Mapping of cosmids

Two µl of cosmid DNA (600 ng) was digested with 3 U of cloning enzyme at 37°C for 3 hours and run on a 0.8% TAE agarose gel. The restriction patterns of the positive cosmids were compared. By looking at the shared bands between the cosmids, a rough map was constructed. One or two overlapping cosmids were then chosen to prepare a plasmid library.

V.C. Verification of the integrity of chosen cosmids

To verify that the cosmid selected did not contain deletions, a Southern hybridization of genomic DNA using the cosmid as a probe was performed. Genomic DNA (3 µg) and the chosen cosmid(s) (50 ng) were digested with 5 U of cloning enzyme at 37°C for 3 hours and loaded side by side on a 0.8% TAE agarose gel. The gel was transferred onto a GeneScreen membrane (Dupont / NEN) using a Tyler VT-20 vacuum transfer unit, by submerging the gel 3 minutes with the depurination solution (0.25 M HCl), 3 minutes with the neutralizing solution (1.5 M NaCl and 0.5 M NaOH) and 30 minutes with the transfer solution (0.4 M NaOH). The blot was rinsed in 2X SSC and irradiated with ultraviolet light for 5 minutes to fix the DNA onto the membrane.

The chosen cosmid was linearized prior probe preparation. Three hundred μg of cosmid was partially digested with 0.5 U of *Hind* III at 37°C for 10 minutes. The Southern hybridization was performed at 38°C following the same procedure used for the colony hybridization. On the autoradiogram, the bands lighting up for the genomic DNA and the digested cosmid should be of the same size.

VI. Preparation of plasmid libraries

VI.A. Nebulization of cosmid

The selected cosmid was sheared into small random fragments. Four μg of cosmid DNA in 500 μl of 25% v/v glycerol was transferred into a modified Hospitak "Up-Mist" medication nebulizer. The DNA was nebulized using nitrogen at 7 p.s.i. for 12 to 15 seconds. The DNA was transferred to a 1.5 ml tube and ethanol precipitated with 1/10 volume of 3 M sodium acetate pH 5.5 and 2 volumes of 95% ethanol. The DNA was washed with 70% ethanol, air dried and dissolved in 20 μl of TE. Five μl of nebulized DNA was incubated with 1 μl of 0.1 μg/μl RNase for 15 minutes at room temperature and then run on a 1% TAE agarose gel at 25V for 18 hours. The size of fragments should be in the range of 1 to 5 kb.

VI.B. End repair of DNA fragments

Two μl of nebulized DNA (0.4 μg) were added to 2 μl of 10X Klenow buffer, 1.5 μl of 1 mM dNTP, 3 U of unmodified T7 DNA polymerase and 2.5 U

of Klenow polymerase (sequencing grade), bringing the final volume to 20 µl with water. The ends of fragments were repaired at 12°C for 30 minutes. The enzymes were denatured when incubated at 65°C for 5 minutes. Ten µl of water, 3 µl of 3 M sodium acetate pH 5.5 and 70 µl of 95% ethanol were added to precipitate the DNA. The fragments were washed in 70% ethanol, air dried and resuspended in 13.6 µl of water.

The DNA fragments were phosphorylated by adding 2 µl of 10 mM ATP, 2 µl 10X PNK buffer and 2.4 µl (24 U) of polynucleotide kinase and incubating at 37 °C for 30 minutes. The enzyme was deactivated when incubated at 65°C for 10 minutes.

VI.C. Ligation

Ten µl of end-repaired DNA was added to 1 µl (50 ng) of *Sma* I - cut and dephosphorylated pUC 18 (Pharmacia), 1 µl of 10X ligase buffer without ATP, 0.3 µl of 10 mM ATP, 1.2 µl of water and 1.5 µl (1.5 U) of T4 DNA ligase. The ligation was done at 12 °C for 18 hours.

VI.D. Transformation

A tube of Epicurian coli XL1-Blue MRF' supercompetent cells (Stratagene) was thawed on ice and 80 µl of cells was transferred into a 14 ml polypropylene tube. The cells were incubated on ice with 1.4 µl of 1:10 β-mercaptoethanol for 10 minutes, mixing gently every 2 minutes. Five µl of ligation was added and the

cells were incubated on ice for 30 minutes. The cells were heat shocked at 42°C for 45 seconds and incubated on ice for 2 minutes. Nine hundred µl of prewarmed SOC broth was added and the cells were grown for 1 hour at 37°C with shaking. One hundred µl and 50 µl amounts were spread onto YT-amp-Xgal-IPTG plates and incubated 18 hours at 37°C.

VI.E. Storage of plasmid libraries

White colonies (300-400) were picked using sterile toothpicks and grown in 100 µl YT-amp broth in 96-well lidded microtiter dishes at 37°C for 18h. The cultures were frozen at -80°C with 15% v/v glycerol.

VII. Screening of plasmid libraries

VII.A. Preparation of blots

Before freezing the microtiter dishes at -80°C, a 96-pronged device was used to stamp the cultures onto GeneScreen Plus membranes (Dupont / NEN) overlaid on top of YT-amp agar in large square dishes. Each microtiter dish was stamped 3 times to obtain triplicates of the plasmid libraries. The plates were incubated at 37°C for 18 hours. The blots were treated, autoclaved, prewashed and scrubbed using the same procedures used for the cosmid libraries.

VII.B. Hybridization of the plasmid libraries

The two probes used for the hybridization of plasmid libraries are Lorist (the section of Tropist that gets packaged into the phage particle) and the

cosmid(s) used to prepare the plasmid library. Prior to the hybridization, Lorist and the cosmid(s) were linearized: 25 ng of Lorist and 300 ng of cosmid(s) were partially digested with 0.5 U of *Hind* III for 10 minutes at 37°C in a final volume of 16 µl. Hybridization of the plasmid libraries was performed at 37°C using the same procedure used for the cosmid libraries. The Lorist probe was used to hybridize one of the three copies of the plasmid libraries, and the cosmid probe was used to hybridize the two other copies of the plasmid libraries.

VIII. Preparation of templates

VIII.A. Nucleobond DNA preparation

On the autoradiogram, plasmids that were positive using the cosmid probe and at the same time, negative using the Lorist probe, were selected to be sequenced. Five ml of YT-amp broth were inoculated directly from the frozen cultures. The cultures were grown with shaking at 37°C for 18 hours.

The plasmids were extracted using the Nucleobond AX PC-20 kit (Macherey/Nagel). The cells were spun 7 minutes and resuspended in 500 µl of buffer S1 (50 mM Tris·Cl, 10 mM EDTA and 400 µg RNase A/ml; pH 8.0). The cells were lysed by adding 500 µl of buffer S2 (200 mM NaOH and 1% SDS) and neutralized by adding 500 µl of buffer S3 (2.8 M KAc; pH 5.2). The neutralized cell lysate was centrifuged for 10 minutes and the supernatant was transferred into a Nucleobond AX cartridge that was equilibrated with 1 ml of buffer N2 (100 mM Tris/H_3PO_4, 15% ethanol and 900 mM KCl; pH 6.3). The cartridges were washed

3 times with 1 ml of buffer N3 (100 mM Tris/H_3PO_4, 15% ethanol and 1150 mM KCl; pH 6.3). The plasmids were eluted with 800 µl of buffer N5 (100 mM Tris/H_3PO_4, 15% ethanol and 1 M KCl; pH 8.5). The purified plasmids were finally precipitated with 600 µl of isopropanol, washed with 70 % ethanol and resuspended in 20 µl of water.

VIII.B. Size screening

The plasmids (0.5 µl) were digested with 3 U of *Pst* I for 3 hours at 37°C and then run on a 1% TAE agarose gel. The sizes of the inserts were recorded. Plasmids with inserts greater than 1 kb were selected to be sequenced. The optical density and the purity of these plasmids were determined using a GeneQuant spectrophotometer.

IX. Sequencing of templates

The sequencing of plasmids was performed using an Applied Biosystems PRISM Dye terminator cycle sequencing core kit by Perkin Elmer. Five hundred ng of template DNA, 3.2 pmole of M13 universal forward or reverse sequencing primer and 7 µl of reaction terminator mix were mixed in a PCR tube with water to obtain a final volume of 20 µl. The cycle sequencing on the Gene Amp PCR system 9600 was set as follows: 96°C for 1.4 minutes; 25 cycles of: 96°C for 10 seconds, 50°C for 5 seconds and 60°C for 4 minutes; and finally hold at 4°C.

Unincorporated terminators were removed by precipitating the sequencing reactions with 2 µl of 3 M of sodium acetate, pH 4.6 and 50 µl of 95% ethanol. The sequencing reactions were incubated on ice for 15 minutes, centrifuged for 20 minutes, washed with 200 µl of 70% ethanol, centrifuged again for 20 minutes and vacuum dried. The pellets were frozen at -20°C until ready to be loaded on the sequencing gel. The sequencing gel was prepared by André Bergeron (University of Ottawa) and run on an Applied Biosystems ABI373 sequencer. The trace data were transferred to a Sun SPARCstation LX for analysis.

X. Construction of contigs maps

The sequences were assembled into contigs and edited using the Staden package, version 1995a (Dear and Staden, 1991). Maps showing the order of the contigs and the distance between these contigs were constructed. The order of contigs can be determined since the orientation of the forward and reverse sequences of a plasmid must converge. The distance between contigs can be determined since the plasmid insert sizes are known.

XI. Joining unlinked contigs

XI.A. Linking contigs via hybridization

In order to construct a contig map, it is necessary to link all the contigs together. When a map is composed of only a few unlinked pieces, additional plasmids in those unlinked regions can be identified through hybridization. In

other words, the plasmid at the end of an unlinked fragment may be used as a probe to identify more plasmids in that particular region.

In order to use a plasmid as a probe, the insert has to be released from the vector. The insert was released by digesting 0.5 µg of plasmid with 3 U of *Bam* HI and *Sac* I or, with 3 U of *Bam* HI and *Eco* RV at 37°C for 3 hours. The digests were run on a 1% TAE agarose gel. The insert released from the double digest, was cut out from the gel. The DNA was extracted from the gel using the GENECLEAN II kit (Bio 101). Briefly, the gel band(s) were melted by adding three volumes of NaI and incubating at 55°C for 10 minutes. Five µl of glassmilk was added and the DNA was incubated at room temperature for 15 minutes, mixing every 2 minutes. The DNA bound to the glassmilk was spun for 10 seconds and the pellet formed was washed three times with 500 µl of New Wash solution (NaCl/ethanol/water). The DNA was dissolved by adding 20 µl of water and incubating at 45°C for 15 minutes. The glassmilk was spun down for 10 seconds and the supernatant containing the DNA was transferred to another tube. Sixteen µl of the gene-cleaned inserts were used to hybridize the plasmid library using the same hybridization protocol as for the plasmid libraries.

XI.B. Linking contigs via PCR amplification

In the case of *M. sedula*, sequencing plasmids in the unlinked regions was not sufficient to link all of the contigs together. The map of *M. sedula* was in fact composed of 2 unlinked pieces. In order to determine the orientation of those 2

unlinked pieces and the distance between them, primers at the end of those unlinked fragments were designed to amplify the gap. Three primers were chosen: the first and second primers 5'—AGAAGGGGTCAGTTGACGTTTC—3' and 5'—TCCATCTCCAGCTTCA CCCTTC—3' were located at the two extremities of the first unlinked fragment and the third primer 5'—CAGGAGAAACTGAACCTTCCAC—3' was located at the end of the second unlinked fragment. A fourth primer at the other extremity of the second unlinked fragment was not chosen since that end contained Tropist3 sequences. The cosmid chosen to prepare *M. sedula*'s plasmid library was used as template DNA for PCR amplification of the gap, using different combinations of primers: the first and second primers and, the first and third primers respectively. One of the 2 combinations will give a PCR product indicating the orientation of the 2 fragments and the length of the gap. The PCR reaction was performed using 10 and 100 ng of cosmid DNA, 4 mM $MgCl_2$, 200 μM dNTP, 30 pmol of each primer, 1X Thermo buffer and 1 U of Taq polymerase. The thermal cycling using a Gene Amp PCR system 9600 was set as follows: 95°C for 5 minutes; 30 cycles of: 1 minute at 94°C, 2 minutes at 65°C and 5 minutes at 72°C; and finally hold at 4°C. Half of the PCR reaction was run on a 1% TAE agarose gel. The primers were further used to sequence part of the gap using the cosmid as the template DNA.

XII. Construction of a phylogenetic tree for the *Sulfolobales* based on the 16S rRNA gene

The 16S rRNA gene sequence from *Methanococcus jannachii* (M59126) *Desulfurolobus mobilis* (M36474), *Acidianus infernus* (X89852), *Acidianus ambivalens* (X90484), *Metallosphaera sedula* (X90481), *Stygiolobus azoricus* (D85520), *Sulfolobus hakonensis* (D14052), *Sulfolobus metallicus* (X90479), *Sulfolobus acidocaldarius* (D14053), *Sulfolobus shibatae* (M32504) and *Sulfolobus solfataricus* P1 (X90478) were retrieved from GenBank. The 16S rRNA gene sequence from *Sulfolobus solfataricus* P2 was obtained from the *Sulfolobus solfataricus* genome project. The 16S rRNA gene sequence from *Sulfolobus tokodaii* and *Sulfolobus solfataricus* MT4 was obtained from this study, from the sequencing of plasmids and also from the sequencing of cosmids using the two primers used to amplify the 16S rRNA fragment.

The sequences were aligned using CLUSTAL W (Thompson *et al.*, 1994). The alignment was not problematic because the 16S rRNA within these archaea is highly conserved in size and sequence, and very few gaps were introduced. The program MEGA (Kumar *et al.*, 1993) was used to determine distances between the sequences assuming the Kimura two-parameter model of sequence evolution. This analysis was performed removing gap sites from the subset data.

A phylogenetic tree was constructed using the phylogeny inference (PHYLIP) software package version 3.5 (Felsenstein, 1993) run on a Sun workstation. A DNA distance matrix was calculated using DNADIST under

maximum likelihood's model for nucleotide substitution. Neighbor-joining trees were constructed using NEIGHBOR. Bootstrap confirmation levels were determined using SEQBOOT to generate 100 data sets and CONSENSE to obtain a majority-rule consensus tree. In all programs, the input of the order of sequences was randomized, jumbled once and the outgroup was specified.

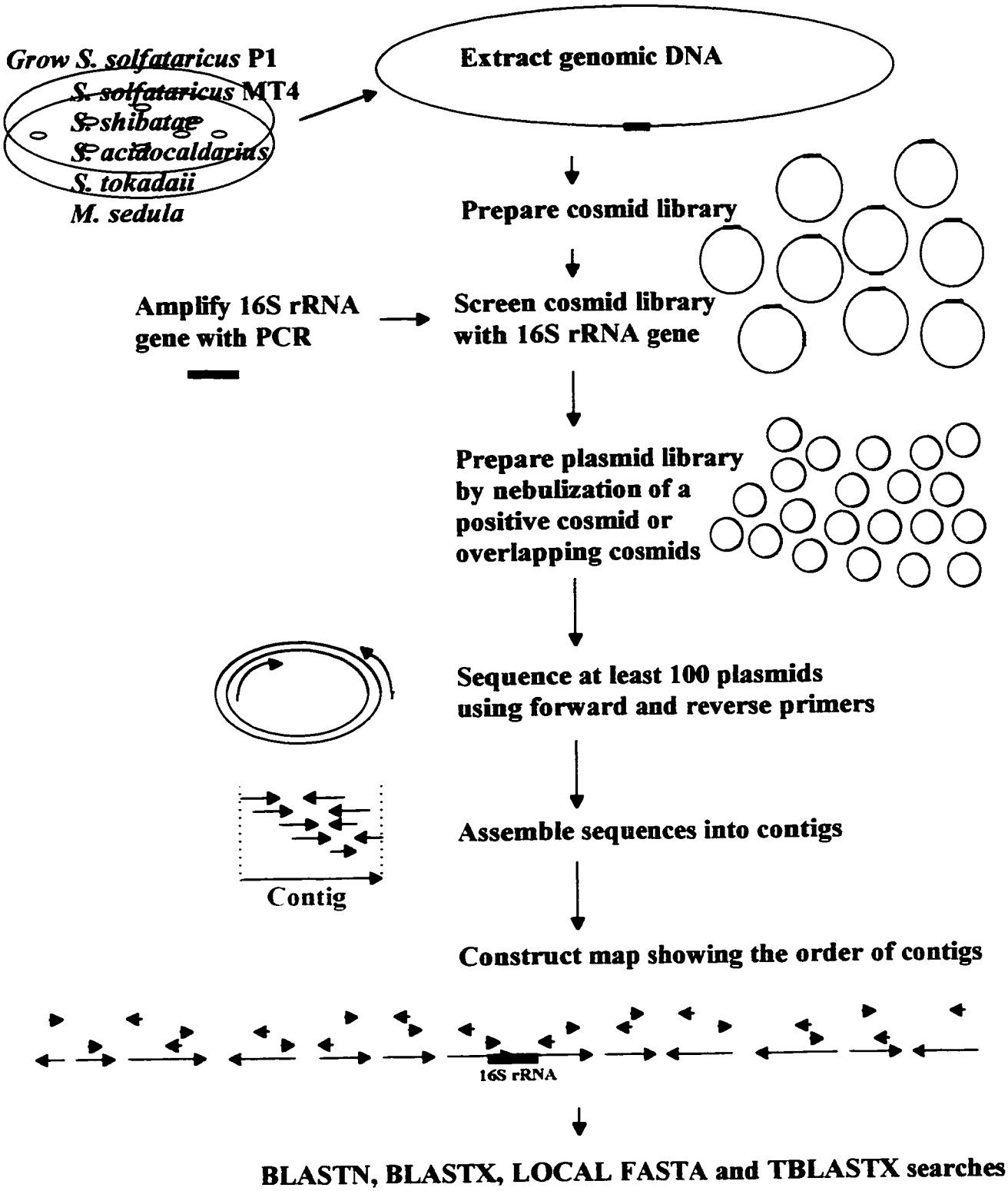
RESULTS

I. Cloning the 16S rRNA gene region of *Sulfolobales*

My research involved comparing gene organization in the 16S rRNA gene region of seven species and strains of *Sulfolobales*. In order to isolate that specific region from the different organisms, cosmid libraries were constructed. Cosmids are useful for this study since they contain large inserts of 40-50 kb. Thus, a single cosmid or two overlapping cosmids that contain the 16S rRNA gene can be chosen to represent the region compared between the different *Sulfolobales*. Figure 4 shows a brief overview of the main steps involved in cloning and sequencing the 16S rRNA gene regions of *Sulfolobales*.

S. solfataricus P1, *S. solfataricus* MT4, *S. shibatae*, *S. acidocaldarius*, *S. tokodaii* and *M. sedula* were first grown in their optimal growth conditions and their genomic DNA was extracted. Digests of genomic DNA confirmed that there was no contamination between the different cultures. Cosmid libraries were then prepared for the different organisms using partially digested genomic DNA. *Xba* I was the cloning enzyme chosen for *S. solfataricus* MT4 whereas *Hin*d III was chosen for *S. solfataricus* P1, *S. shibatae*, *S. acidocaldarius*, *S. tokodaii* and *M. sedula*. The cosmid libraries were screened with the 16S rRNA gene, which was amplified by PCR from each organism. Southerns of genomic DNA have shown that the 16S rRNA gene of the different *Sulfolobales* studied was present as a single-copy gene (Kondo *et al.*, 1993, and data not shown).

Figure 4. Overview of the main steps involved in comparing the 40-50 kbp 16S rRNA gene regions of different *Sulfolobales*.


Grow S. solfataricus P1
S. solfataricus MT4
S. shibatae
S. acidocaldarius
S. tokadaii
M. sedula
Extract genomic DNA
Prepare cosmid library
Amplify 16S rRNA gene with PCR
Screen cosmid library with 16S rRNA gene
Prepare plasmid library by nebulization of a positive cosmid or overlapping cosmids
Sequence at least 100 plasmids using forward and reverse primers
Contig
Assemble sequences into contigs
Construct map showing the order of contigs
16S rRNA
BLASTN, BLASTX, LOCAL FASTA and TBLASTX searches

Results from the colony hybridizations of the cosmid libraries, showed that 10 cosmids out of 576 were found positive for *S. solfataricus* P1, 7 cosmids out of 1152 were found positive for *S. solfataricus* MT4, 7 cosmids out of 576 were found positive for *S. shibatae*, only one cosmid out of 960 was found positive for *S. acidocaldarius*, 14 cosmids out of 288 were found positive for *S. tokodaii*, and 6 cosmids out of 619 were found positive for *M. sedula*. The restriction patterns of these positive cosmids were compared. By looking at the shared bands between the different positive cosmids, rough maps showing how cosmids overlapped were constructed (Figure 5). Based on these maps, one or two overlapping cosmids were chosen to represent the 16S rRNA gene region for each species. Cosmids P1h02g05 and P1h04c08 which covered a 52.5 kb region were chosen for *S. solfataricus* P1; cosmids MT4x05c09 and MT4x05f09 which covered a 48.5 kb region were chosen for *S. solfataricus* MT4; a single 44 kb cosmid, shih03b12 was chosen for *S. shibatae*; a single 37.5 kb cosmid, Sah07d09 was chosen for *S. acidocaldarius*; cosmids tokh01f01 and tokh03b07 which covered a 46.5 kb region were chosen for *S. tokodaii*; and cosmids mh03h09 and mh06e08 which covered a 59 kb region were chosen for *M. sedula*. To determine if the selected cosmids carried deletions, the different cosmids were used as probes against genomic digests. Results from the Southern hybridizations of *S. solfataricus* P2, *S. shibatae*, *S. acidocaldarius* and *M. sedula* confirmed that the cosmids did not carry deletions. Figure 6A is an example of the Southern hybridization of

Figure 5. Maps of cosmids that are overlapping in the 16S rRNA gene regions of *M. sedula* (A), *S. tokodaii* (B), *S. acidocaldarius* (C), *S. shibatae* (D), *S. solfataricus* MT4 (E) and *S. solfataricus* P1 (F). The cosmids that were chosen to represent the 16S rRNA gene regions in my comparative analysis are shown in bold.

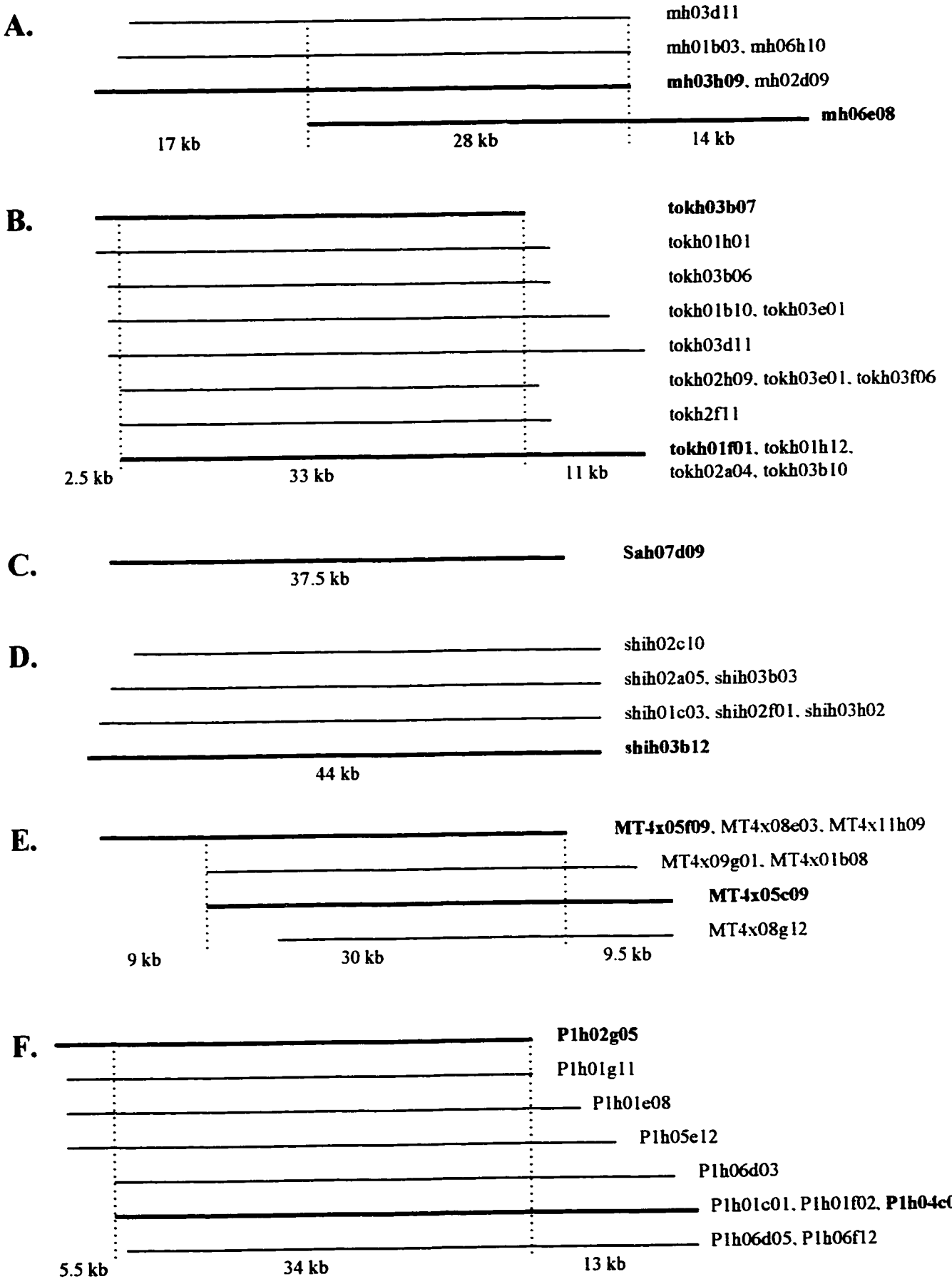

A.
mh03d11
mh01b03, mh06h10
mh03h09, mh02d09
mh06e08
17 kb
28 kb
14 kb
B.
tokh03b07
tokh01h01
tokh03b06
tokh01b10, tokh03e01
tokh03d11
tokh02h09, tokh03e01, tokh03f06
tokh2f11
tokh01f01, tokh01h12,
tokh02a04, tokh03b10
2.5 kb
33 kb
11 kb
C.
Sah07d09
37.5 kb
D.
shih02c10
shih02a05, shih03b03
shih01c03, shih02f01, shih03h02
shih03b12
44 kb
E.
MT4x05f09, MT4x08e03, MT4x11h09
MT4x09g01, MT4x01b08
MT4x05c09
MT4x08g12
9 kb
30 kb
9.5 kb
F.
P1h02g05
P1h01g11
P1h01e08
P1h05e12
P1h06d03
P1h01c01, P1h01f02, P1h04c08
P1h06d05, P1h06f12
5.5 kb
34 kb
13 kb

Figure 6. Southern hybridizations of genomic digests using as a probe cosmid shih03b12 for *S. shibatae* (A), cosmid tokh03b07 for *S. tokodaii* (B) and cosmid mt4x05f09 for *S. solfataricus* MT4 (C). In (A), genomic DNA and cosmid shih03b12 from *S. shibatae* were digested with *Hin*d III and hybridized with cosmid shih03b12. The hybridization confirms that cosmid shih03b12 did not carry deletions. The decrease in relative intensity at the top of the gel can probably be explained by a poor transfer of high molecular weight fragments. In (B), genomic DNA and cosmid tokh03b07 from *S. tokodaii* were digested with *Hin*d III and hybridized with cosmid tokh03b07. The presence of extra bands suggests that cosmid tokh03b07 carries a repeated element. The differences in relative intensities between cosmid and genomic lanes can probably be explained by an increase of contrast in the scan of the autoradiogram relative to the original. In (C), genomic DNA and cosmid mt4x05f09 from *S. solfataricus* MT4 were digested with *Xba* I and hybridized with cosmid mt4x05f09. The presence of extra bands suggests that cosmid mt4x05f09 also carries a repeated element. In addition, maps of cosmids mt4x05f09 and mt4x05c09 have shown some differences in the overlapping region due to transposition of insertion elements. The genomic DNA of *S. solfataricus* MT4 was therefore not homogeneous. The white arrows indicate the location of Tropist vector, present in the cosmid digests and absent in the genomic digests. In (A), the cosmid has an additional band of the same size as the vector. Bands in the marker lane are also lighting up since Tropist is based on lambda, and many of these bands are partial digests of lambda.

marker
genomic DNA from S. shibatae digested with Hind III
cosmid shih03b12 digested with Hind III
marker
genomic DNA from S. tokodaii digested with Hind III
cosmid tokh03b07 digested with Hind III
marker
cosmid mt4x05f09 digested with Xba I
genomic DNA from S. solfataricus MT4 digested with Xba I
A
B
C

S. *shibatae*; the bands lighting up for the *Hin*d III-digested genomic DNA and cosmid shih03b12, are identical when hybridized with cosmid shih03b12. On the other hand, results from the Southern hybridizations of *S. tokodaii* and *S. solfataricus* MT4, detected many extra bands in addition to the restriction fragments of the cosmids (Figure 6B and 6C). These additional bands suggest the presence of repeated sequences in these cosmids.

II. Sequencing the 16S rRNA region of *Sulfolobales*

Plasmid libraries were constructed for the different *Sulfolobales* by nebulizing one or two overlapping cosmids and, cloning the inserts into a plasmid vector. One plasmid library was constructed for *S. solfataricus* P1 using cosmids P1h02g05 and P1h04c08 combined; two plasmid libraries were built for *S. solfataricus* MT4 using cosmids MT4x05c09 and MT4x05f09 separately; one plasmid library was prepared for *S. shibatae* using cosmid shi03b12; one plasmid library was built for *S. acidocaldarius* using cosmid Sah07d09; two plasmid libraries were constructed for *S. tokodaii* using cosmids tokh01f01 and tokh03b07 separately; and one plasmid library was prepared for *M. sedula* using cosmids mh03h09 and mh06e08 combined. The plasmids were randomly sequenced using the universal forward and reverse primers. More specifically, 24 plasmids were sequenced for *S. solfataricus* P1, 112 plasmids were sequenced for *S. solfataricus* MT4; 105 plasmids were sequenced for *S. shibatae*; 100 plasmids were sequenced for *S. acidocaldarius*; 115 plasmids were sequenced for *S. tokodaii*; and 160

plasmids were sequenced for *M. sedula*. The average length of sequences was approximately 500 nt.

III. The assembly of sequences into contig maps

The different sequences were assembled into contigs since many sequences overlapped the same region. A contig is in fact the consensus of overlapping sequences. Maps of contigs were constructed for *S. solfataricus* MT4, *S. shibatae*, *S. acidocaldarius*, *S. tokodaii* and *M. sedula*. The order and orientation of contigs relative to each other, can be determined since the forward and reverse sequences of a plasmid converge. The distances between the different contigs can be determined since plasmid insert sizes are known. It was imperative that all contigs for each species be linked to each other in a map. When necessary, more plasmids in the unlinked regions were sequenced until all contigs were connected in a map. In *M. sedula*'s case, in order to connect two unlinked regions, primers at the ends of these unlinked regions were designed to amplify and partly sequence the gap. The different maps built were composed of 30 contigs for *S. solfataricus* MT4, 25 contigs for *S. shibatae*, 23 contigs for *S. acidocaldarius*, 28 contigs for *S. tokodaii* and 19 contigs for *M. sedula*. The different contigs varied in size, from 225 bp to 11 832 bp. Appendix II is a table containing all contig names, length of contigs, orientation and position of contigs on the maps for *S. solfataricus* MT4, *S. shibatae*, *S. acidocaldarius*, *S. tokodaii* and *M. sedula*.

A contig map was not constructed for *S. solfataricus* P1 since the 16S rRNA region of *S. solfataricus* P1 was found to be identical to that of *S. solfataricus* P2. Cosmids sh22e04.19 and sh04g11.23 from *S. solfataricus* P2 are two overlapping cosmids that contain the 16S rRNA gene. The identical restriction patterns observed in cosmid P1h02g05 from *S. solfataricus* P1 and cosmid sh22e04.19 from *S. solfataricus* P2 was the first indication that these regions had the same organization. In addition, when 24 plasmids were sequenced using the universal forward and reverse primers, the sequences all matched cosmid sh22e04.19 and sh04g11.23 from *S. solfataricus* P2. When these sequences were mapped onto *S. solfataricus* P2, the sizes of the insert of plasmids sequenced were consistent with the organization of the cosmid maps of sh22e04.19 and sh04g11.23.

IV. Similarity searches performed on the contig sequences

All contig sequences generated from this study, in addition to the sequences available from the genome project of *S. solfataricus* P2, were used for comparative sequence analyses. The contig sequences were first subjected to a number of BLAST (BLASTN, BLASTX and TBLASTX) (Pearson and Lipman, 1988) and FASTA (LOCAL FASTA) (Altschul *et al.*, 1990) searches. When performing these searches, only scores lower than e^{-5} for BLAST and scores higher than 400 for FASTA searches were considered. The probability that these similarity matches are real increases as BLAST scores decrease and FASTA scores increase.

To identify the genes present in the 16S rRNA gene regions, the contig sequences were compared against the NCBI database. BLASTN and BLASTX searches were performed at the nucleotide and amino acid level, respectively. BLASTX is more sensitive in finding similarity matches, yet BLASTN searches are useful in identifying the 16S and 23S rRNA genes, and insertion sequences. Results from these searches are shown in Appendix III (A-E).

The contig sequences were then compared to sequences from *S. solfataricus* P2. To accomplish these comparisons, I constructed a database containing all the sequences of cosmids from the genome sequencing project of *S. solfataricus* P2, that were available in November of 1997. LOCAL FASTA and TBLASTX are the searches performed on the contig sequences at the nucleotide and amino acid level, respectively. Results from these searches are shown in Appendix IV (A-F). When contigs share similarities with cosmids sh22e04.19 (c19) and sh04g11.23 (c23), it could suggest that some conservation in the gene organization of the 16S rRNA region exists. Appendix IV (F) shows the results obtained from LOCAL FASTA and TBLASTX searches performed on the contig sequences of *S. solfataricus* P1. All sequences from *S. solfataricus* P1 matched cosmids sh22e04.19 (c19) and sh04g11.23 (c23) from *S. solfataricus* P2.

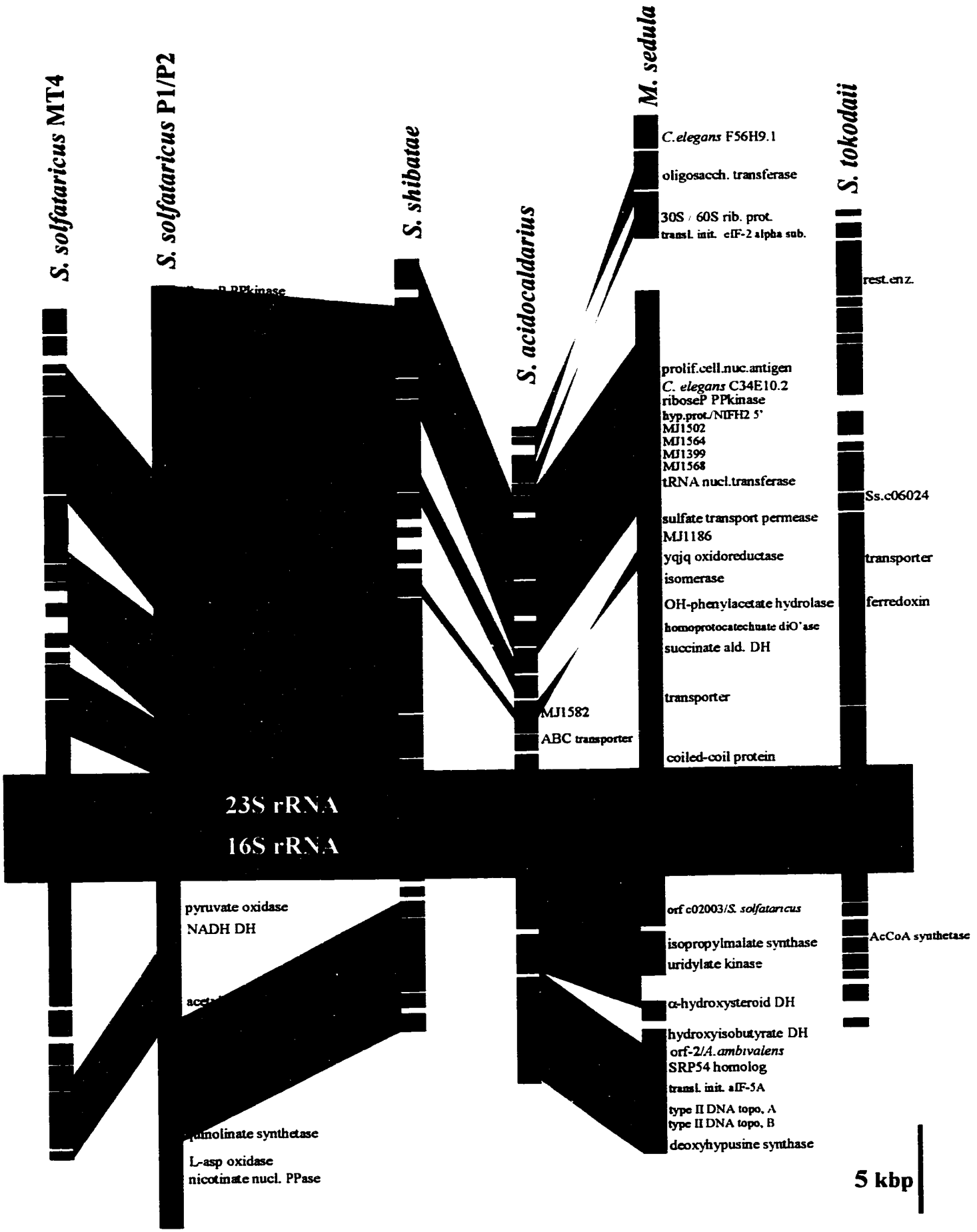
Finally, each contig sequence was compared against all contig sequences from every *Sulfolobales* studied. To perform this comparison, a database containing all sequences of contigs was constructed. TBLASTX searches at the

amino acid level only were performed. Results from these searches are shown in Appendix V (A-E).

V. Comparative sequence analyses

The contig maps of the different *Sulfolobales* are shown in Figure 7. Each pink block represents the sequence of a contig. The contig maps of *S. solfataricus* P1 and *S. solfataricus* P2 are represented by one large contig, while the contig maps of *S. solfataricus* MT4, *S. shibatae*, *S. acidocaldarius*, *M. sedula* and *S. tokodaii* are represented by many contigs. All contig maps have been aligned at the level of the 16S and 23S rRNA genes. The different scores obtained from the similarity searches were examined, in order to find blocks of conserved gene organization. The cyan zones in Figure 7 represent regions of similarity in gene organization between organisms. *S. solfataricus* MT4 shares many similarities with *S. solfataricus* P2. The differences observed are mainly due to the presence of repeated sequences in *S. solfataricus* MT4. Conservation in gene organization is even more apparent when comparing *S. solfataricus* P2 and *S. shibatae*. The only difference in gene organization between these two organisms is the region upstream from the 16S rRNA gene. Many conserved regions were also observed when comparing *S. acidocaldarius* and *M. sedula*. Three blocks of 6, 9 and 12 kbp are conserved in their gene organization. Moreover, one of these conserved blocks is also present in *S. shibatae* and *S. solfataricus* P2 in the same organization. *S. acidocaldarius* and *M. sedula* share a few other small regions of

Figure 7. Comparative sequence analysis of the 16S rRNA gene region of seven *Sulfolobales*. The different contig maps are shown. Each pink block represents the sequence of a contig. Insertion sequences are shown in green. All contig maps have been aligned at the level of the rRNA operon. BLAST and FASTA searches were performed on the contigs to identify similarities in gene organization. (Appendix III, IV and V) The cyan zones represent blocks of conserved gene organization between the organisms. Bar, 5 kbp.

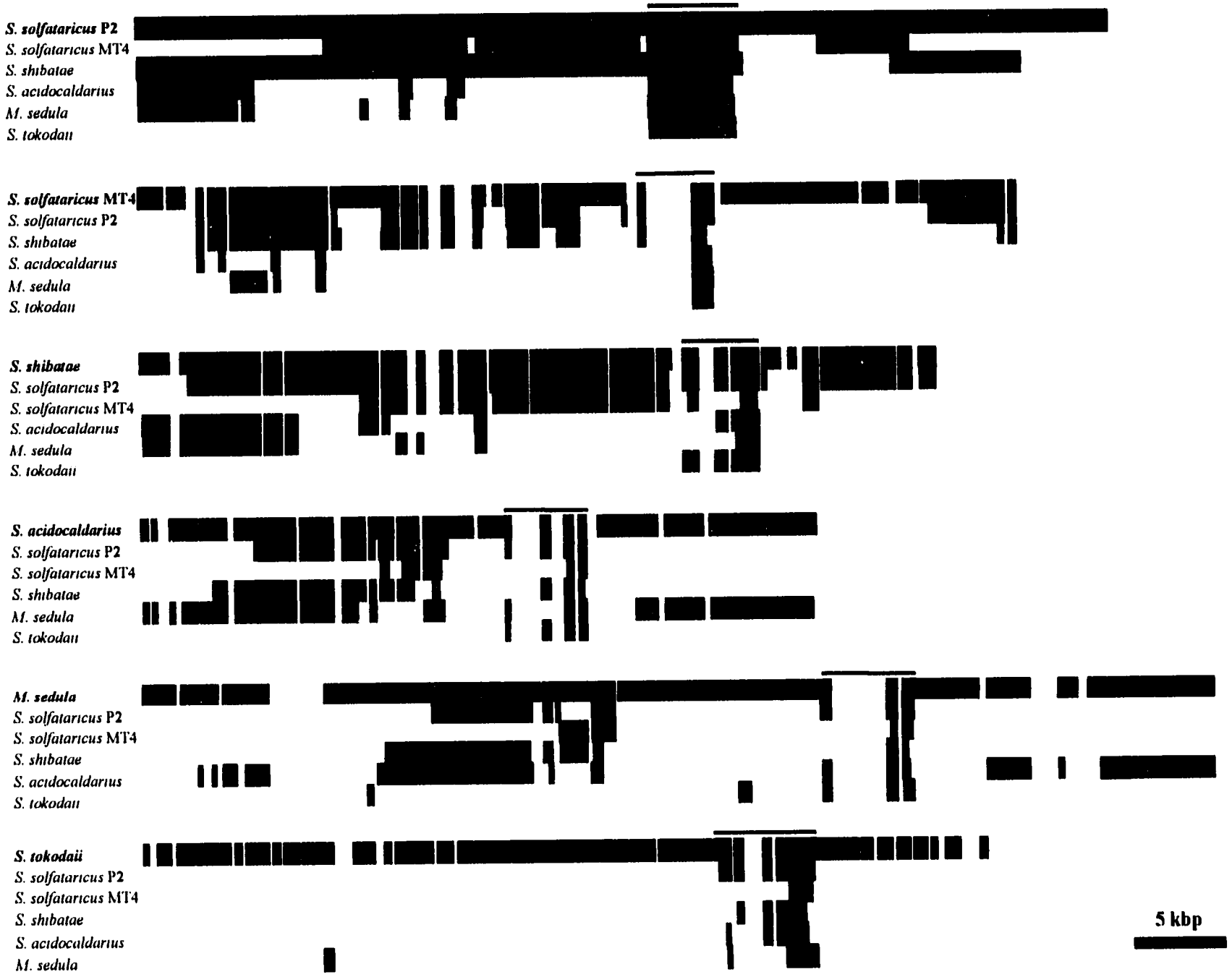

S. solfataricus MT4
S. solfataricus P1/P2
S. shibatae
S. acidocaldarius
M. sedula
S. tokodaii
C.elegans F56H9.1
oligosacch. transferase
30S / 60S rib. prot.
transl. init. eIF-2 alpha sub.
rest.enz.
prolif.cell.nuc.antigen
C. elegans C34E10.2
riboseP PPkinase
hyp.prot./NIFH2 5'
MJ1502
MJ1564
MJ1399
MJ1568
tRNA nucl.transferase
Ss.c06024
sulfate transport permease
MJ1186
yqjq oxidoreductase
transporter
isomerase
ferredoxin
OH-phenylacetate hydrolase
homoprotocatechuate diO'ase
succinate ald. DH
transporter
MJ1582
ABC transporter
coiled-coil protein
23S rRNA
16S rRNA
pyruvate oxidase
NADH DH
orf c02003/S. solfataricus
AcCoA synthetase
isopropylmalate synthase
uridylate kinase
α-hydroxysteroid DH
hydroxyisobutyrate DH
orf-2/A.ambivalens
SRP54 homolog
transl. init. aIF-5A
type II DNA topo. A
type II DNA topo. B
deoxyhypusine synthase
L-asp oxidase
nicotinate nucl. PPase
5 kbp

similarity, but in the opposite orientation. Finally, *S. tokodaii* does not share any similarities with the other *Sulfolobales*, except for the 16S and 23S rRNA genes. *S. tokodaii* has however, many insertion sequences in that particular region of the genome.

Figure 8 is another representation of the similarities observed in gene organization. In this comparative analysis, every contig from a species is compared against total contigs from all other organisms. Thus, six different comparisons were performed. In each comparison, the contig map of the organism being compared is shown on top and any similarities to these are shown in colored blocks. Thus, blank regions indicate that these sequences are not present in that organism or, that the data are not available. Data are missing when the sequence is within a gap or beyond the end of the cosmid. For this reason, it is important to refer to Figure 7 at the same time to avoid any misleading interpretation. Figure 8 together with Figure 7, show all similarities in gene organization between the different *Sulfolobales*. Many similarities can be observed; particularly between *S. solfataricus* and *S. shibatae* and, between *S. acidocaldarius* and *M. sedula*. This comparison is very useful for finding breakpoints, where rearrangements occurred.

Dot plots at the amino acid level were also performed (by R.L. Charlebois) to illustrate further the similarities observed in gene organization between the different *Sulfolobales* (Appendix VI). Each dot plot has a 33 amino acid resolution and, will insert a dot every time 5 amino acids out of 8 are matched.

Figure 8. Comparative sequence analyses of the different *Sulfolobales*. In each comparison, the top organism in bold is compared against the sequences from the other *Sulfolobales*. Pink blocks represent regions of similarity. Dark purple blocks represent regions of similarity, but in the opposite orientation. A blue line indicates the position of the rRNA operon on the contig map. Blank areas can be an indication that the sequence is absent in that organism or, that the data are missing because this region is within a gap or at the end of a cosmid. Bar, 5 kbp.


S. solfataricus P2
S. solfataricus MT4
S. shibatae
S. acidocaldarius
M. sedula
S. tokodaii
S. solfataricus MT4
S. solfataricus P2
S. shibatae
S. acidocaldarius
M. sedula
S. tokodaii
S. shibatae
S. solfataricus P2
S. solfataricus MT4
S. acidocaldarius
M. sedula
S. tokodaii
S. acidocaldarius
S. solfataricus P2
S. solfataricus MT4
S. shibatae
M. sedula
S. tokodaii
M. sedula
S. solfataricus P2
S. solfataricus MT4
S. shibatae
S. acidocaldarius
S. tokodaii
S. tokodaii
S. solfataricus P2
S. solfataricus MT4
S. shibatae
S. acidocaldarius
M. sedula
5 kbp

The intensity of the dot is also indicative of the number of matches that are found within the window of 33 amino acids. If a match is found when one of two sequences is in the opposite orientation, the dot is shown in red instead of black. Gray areas represents regions that have not been sequenced. Thus, if gene organization is conserved between two species, a straight line should be observed. The intensity of the line is also indicative of the relatedness of the two organisms compared. These dot plot comparisons basically confirm the similarities in gene organization shown in Figures 7 and 8. The advantage of dot plots is in being able to detect easily insertions, deletions and replacements of sequences in one of the two organisms compared. For instance, insertions of repeated elements in *S. solfataricus* MT4 can be detected since the different lines are not contiguous. Another example, a large deletion or insertion was detected when comparing *S. solfataricus* P2 to *S. shibatae*. Dot plots are also useful in simply finding blocks of conserved gene organization. For instance, the 7-8 kbp conserved region in *S. solfataricus* P2, *S. shibatae*, *S. acidocaldarius* and *M. sedula* is shown as a straight line. Many conserved blocks are also observed between *M. sedula* and *S. acidocaldarius*. A few genes that have been inverted in *M. sedula* are shown by short red lines. And finally, as expected, the only conservation observed between *S. tokodaii* and the other *Sulfolobales*, is the rRNA operon.

A qualitative tree was constructed showing the different patterns observed when examining the similarities in gene organization between the different

Sulfolobales (Figure 9). *S. solfataricus* P1/P2 and *S. shibatae* are grouped together due to their high conservation in gene organization. *S. solfataricus* MT4 also groups with these two, but not as closely mainly because of the presence of insertion sequences in *S. solfataricus* MT4. *S. acidocaldarius* and *M. sedula* are grouped together because they share many blocks of conserved gene organization. The *S. solfataricus* / *S. shibatae* and *S. acidocaldarius* / *M. sedula* groups are further organized together since they all share a 7-8 kbp conserved block in their gene organization. Finally, *S. tokodaii* is an outgroup since it does not share any similarities in gene organization with the other *Sulfolobales*, except for the rRNA operon.

VI. Comparison of the 16S rRNA gene of *Sulfolobales*

To determine the distances between the 16S rRNA sequences of the different *Sulfolobales* studied, the 16S rRNA gene of *S. solfataricus* MT4 and *S. tokodaii* were sequenced. The 16S rRNA sequence of *S. solfataricus* P2 was obtained from the sequencing genome project, whereas the 16S rRNA gene sequences of *S. solfataricus* P1, *S. shibatae*, *S. acidocaldarius*, *M. sedula* and other archaea were retrieved from GenBank. After aligning the sequences, estimates of substitutions per nucleotide were calculated using Kimura's correction method (Table 1). Figure 10 shows a bootstrapped neighbor-joining tree derived from the distances calculated between the different 16S rRNA gene sequences, assuming maximum likelihood's model of evolution. *Methanococcus voltae*, a

Figure 9. Qualitative tree derived from the similarities observed in gene organization, in the 16S rRNA gene regions of different *Sulfolobales*. The relationships showing the different arrangements in gene organization were determined by a simple visual inspection of Figure 7 and Figure 8.

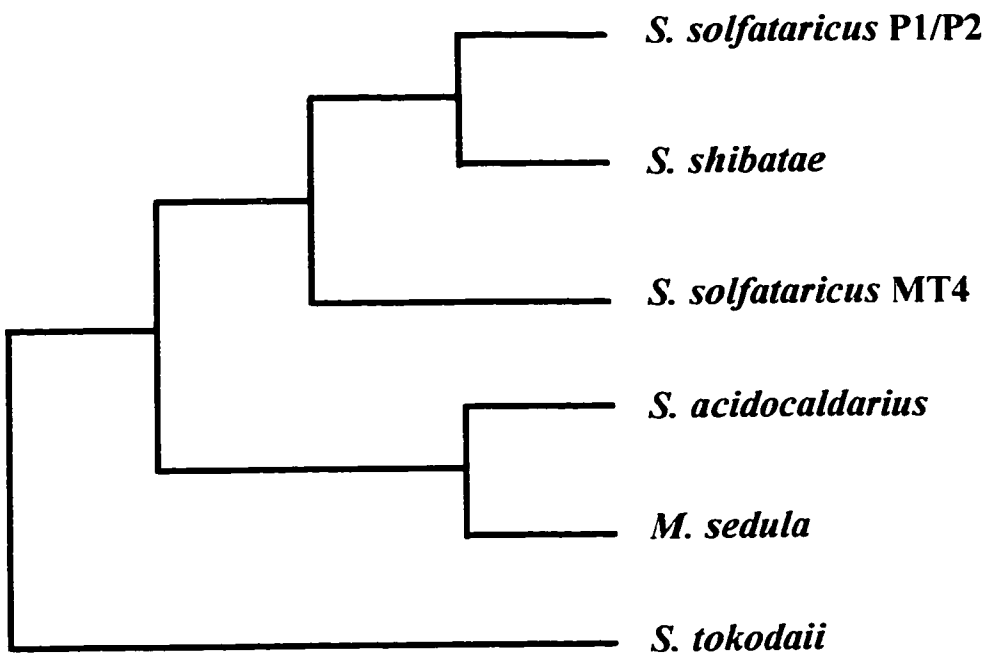
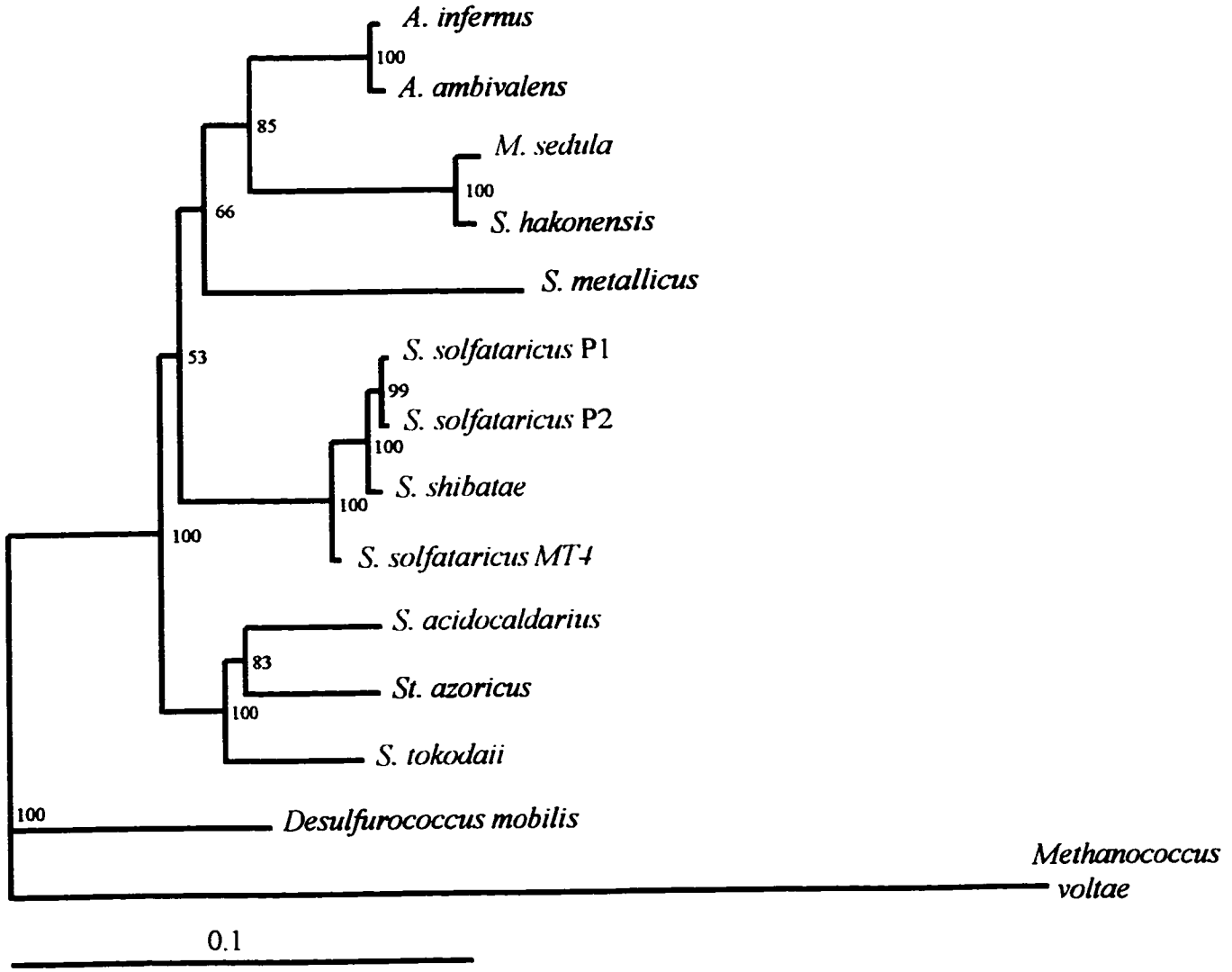

S. solfataricus P1/P2
S. shibatae
S. solfataricus MT4
S. acidocaldarius
M. sedula
S. tokodaii

Table 2. Multiple comparison matrix of estimated nucleotide substitutions per site assuming Kimura's correction method, of seven 16S rRNA gene sequences from *Sulfolobales*. The standard errors are shown in brackets.

	S. solfataricus P2	*S. solfataricus* P1	*S. solfataricus* MT4	*S. shibatae*	*S. acidocaldarius*	*S. tokodaii*	*M. sedula*
***S. solfataricus* P2**		0.0000 (0.0000)	0.0137 (0.0041)	0.0049 (0.0025)	0.0764 (0.0101)	0.0738 (0.0100)	0.1044 (0.0121)
***S. solfataricus* P1**	0.0000 (0.0000)		0.0137 (0.0041)	0.0049 (0.0025)	0.0764 (0.0101)	0.0738 (0.0100)	0.1044 (0.0121)
***S. solfataricus* MT4**	0.0137 (0.0041)	0.0137 (0.0041)		0.0112 (0.0037)	0.0765 (0.0100)	0.0765 (0.0102)	0.1044 (0.0121)
S. shibatae	0.0049 (0.0025)	0.0049 (0.0025)	0.0112 (0.0037)		0.0764 (0.0101)	0.0738 (0.0100)	0.1044 (0.0121)
S. acidocaldarius	0.0764 (0.0101)	0.0764 (0.0101)	0.0765 (0.0100)	0.0764 (0.0101)		0.0561 (0.0086)	0.1057 (0.0121)
S. tokodaii	0.0738 (0.0100)	0.0738 (0.0100)	0.0765 (0.0102)	0.0738 (0.0100)	0.0561 (0.0086)		0.1034 (0.0121)
M. sedula	0.1044 (0.0121)	0.1044 (0.0121)	0.1044 (0.0121)	0.1044 (0.0121)	0.1057 (0.0121)	0.1034 (0.0121)	

Figure 10. Bootstrapped neighbor-joining tree based on the 16S rRNA nucleotide sequences of different archaea, assuming maximum likelihood's model of evolution. The organisms shown in blue are the different *Sulfolobales* used in my comparative sequencing analysis. Bar, 0.1 substitutions / site.


A. infernus
100
A. ambivalens
85
M. sedula
100
S. hakonensis
66
S. metallicus
53
S. solfataricus P1
99
S. solfataricus P2
100
S. shibatae
100
S. solfataricus MT4
100
S. acidocaldarius
83
St. azoricus
100
S. tokodaii
100
Desulfurococcus mobilis
Methanococcus voltae
0.1

euryarchaeote and *Desulfurococcus mobilis*, a crenarchaeote belonging to the order *Thermoproteales* were the outgroups used. The branching of the tree shows that *S. solfataricus* P1, *S. solfataricus* P2, *S. solfataricus* MT4 and *S. shibatae* form one group whereas *S. acidocaldarius* and *S. tokodaii* form another.

DISCUSSION

I. Summary of the major findings

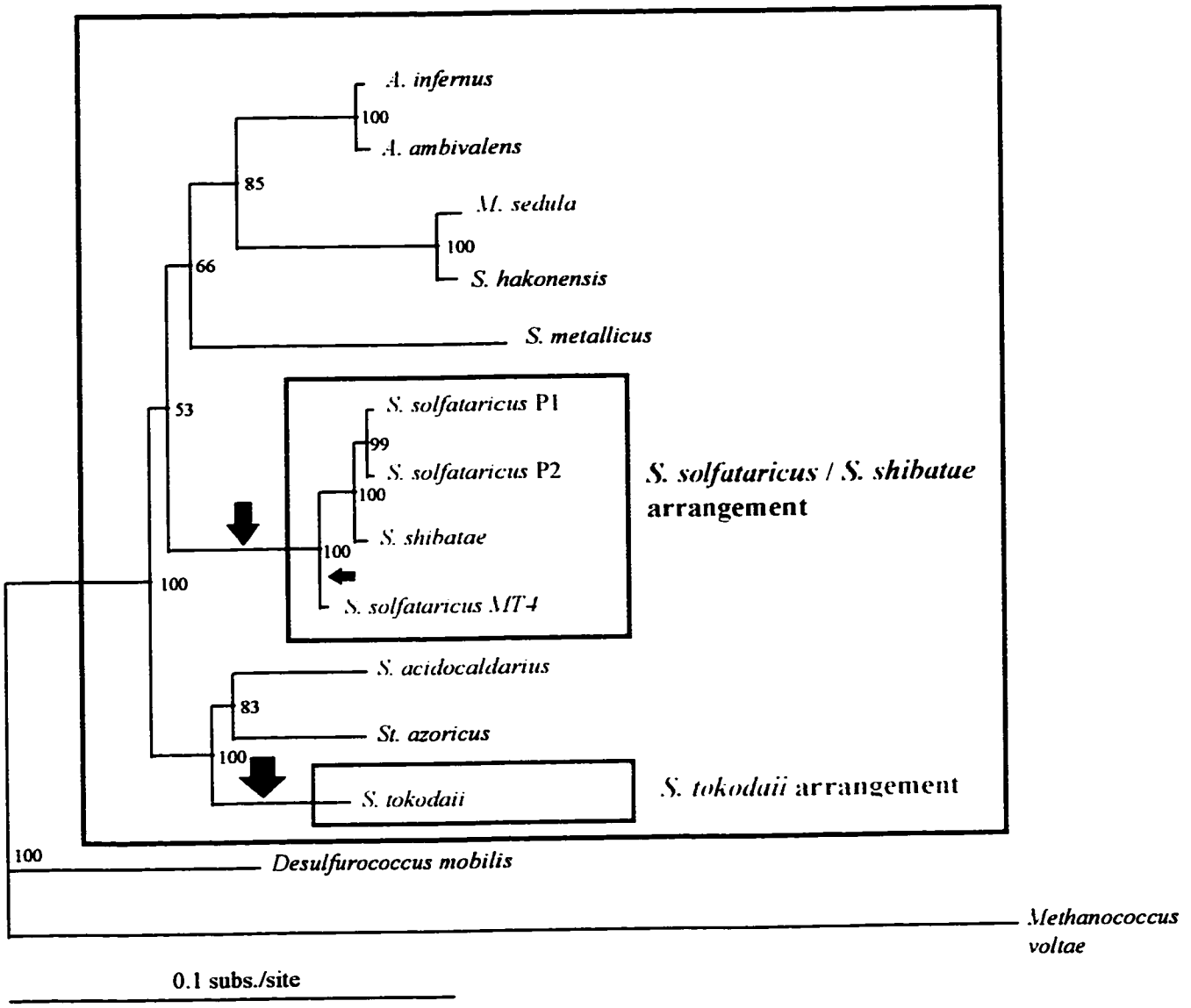
My research involved comparing an equivalent 40-50 kbp region of seven *Sulfolobales* at the sequence level. This comparative analysis has shown that some conservation in gene order exists in that particular region of the genome of *Sulfolobales*. Figure 7 and Figure 8 show that gene organization of *S. solfataricus* P1/P2/MT4 and *S. shibatae* is similar, and that gene organization of *S. acidocaldarius* and *M. sedula* is similar. These six *Sulfolobales* further share a 7-8 kbp block region, where gene organization is conserved. *S. tokodaii* does not share any similarities with the other *Sulfolobales* apart from the ribosomal operon. Based on these similarities in gene organization, a qualitative tree was constructed showing the different organizations observed in the 16S rRNA gene regions (Figure 9). The actual distance tree based upon the 16S rRNA gene sequences was also constructed for comparison with the qualitative tree obtained (Figure 10).

II. Gradual versus punctuated model of evolution

When the qualitative and distance trees obtained were compared, some differences were observed. The qualitative tree differs from the distance tree by grouping *S. acidocaldarius* and *M. sedula* together and, by placing *S. tokodaii* as an outgroup to all the other *Sulfolobales* studied. In Figure 11, I have integrated

Figure 11. The information obtained from the qualitative tree was integrated into the bootstrapped neighbor-joining tree. The different gene arrangements observed in the 16S rRNA regions of the seven *Sulfolobales* studied are shown. I can hypothesize that the ancestor of the *Sulfolobales* had the *S. acidocaldarius* / *M. sedula* gene arrangement. The purple arrows indicate that at some stage, rearrangements in gene organization in that particular region of the genome occurred. These gene shuffling events gave rise to the *S. solfataricus* / *S. shibatae* and *S. tokodaii* gene arrangements.

***S. acidocaldarius* / *M. sedula* arrangement**



the information obtained from the gene organization analysis with the actual distance tree. From this figure, a few hypotheses can be suggested. First, I can hypothesize that the ancestor of the *Sulfolobales* had the *S. acidocaldarius* / *M. sedula* gene arrangement in the 16S rRNA gene region. Extensive rearrangements arose in that particular region of *S. tokodaii*. Rearrangements at some stage occurred to give rise to the *S. solfataricus* / *S. shibatae* arrangement. Other rearrangements also occurred in *S. solfataricus* MT4 without altering too much the *S. solfataricus* / *S. shibatae* arrangement. Both distance and qualitative trees agreed that *S. solfataricus* P1/P2 and *S. shibatae* were more closely related compared to *S. solfataricus* MT4. These observations suggest that *S. solfataricus* is not a monophyletic group. Furthermore, the distance tree shows that the genus *Sulfolobus* is also not a monophyletic group. Thus, the naming of different species and strains within the genus *Sulfolobus* would need to be re-appraised.

The null hypothesis of my research predicts that rearrangements in gene organization mirror phylogenetic history. Thus, the hypothesis implies that rearrangements in gene organization accumulate constantly and gradually over time. The results obtained from my comparative analysis showed that there was no strong correlation between similarities in gene organization in the region surrounding the 16S rRNA gene of 7 *Sulfolobales* and their evolutionary distance based upon the 16S rRNA sequences. My hypothesis was therefore refuted on the basis that my comparative analysis did not support a gradual model of evolution.

Alternatively, my comparison supported a punctuated model of evolution, where genes were shuffled periodically. Shuffling genes around can be advantageous along the evolutionary path, when an organism needs to respond to an environmental stress. Under such stress, the forces disrupting gene order may become stronger than the forces maintaining gene order (Charlebois and St. Jean, 1995). Punctuated evolution, however, may still result from chance events, where one rearrangement triggers others. In Figure 11, purple arrows represent stages where genes were possibly shuffled, to give rise to the different gene organizations observed today.

Some of the differences observed in gene organization could be the result of lateral transfer. Nonetheless, it is unlikely that the 16S/23S rRNA locus itself was moved by horizontal gene transfer, since these rRNAs must participate in macromolecular assemblies with other cellular components. Comparing phylogenies of the genes found in the region surrounding the rRNA genes, to the phylogeny of these rRNA genes could determine if in fact any lateral transfers occurred.

III. Gene order, a promising parameter to build phylogenies?

III.A. Gene evolution versus genome evolution

Gene evolution can infer phylogenetic distances from sequences of individual genes. Until now, nucleic acid sequencing has been the most powerful

and direct method for comparing genomes. Lots of attention has been given to the rRNA genes since their structure changes very slowly in time, due to their constant and important function. To measure the divergence between two organisms, the sequences of a gene are first aligned. Since the alignment does not allow any intersection between pairs of nucleotides, duplications, inversions and transpositions are not reflected in the distances calculated. Only nucleotide replacements, deletions and insertions are considered.

Genome evolution can however infer phylogenetic distances from gene orders, without discarding duplications, inversions and transpositions (Sankoff and Goldstein, 1989). Rearrangements in gene organization can be measured by determining the minimal steps of chromosomal inversions, transpositions, insertions and deletions required to convert the order of one genome to another. Another benefit of genome evolution is that the phylogenetic distances inferred from gene orders are not influenced by the different substitution rates of genes.

III.B. Can gene order be used to build phylogenies?

Whether or not gene order can be used as a parameter to construct phylogenies still remains uncertain. Gene order can be a true indicator of phylogenetic relatedness, only if genomic rearrangements follow a gradual model of evolution. To address this question, phylogenetic trees that were constructed on the basis of gene order must be compared with standard phylogenies. Many comparisons of gene organization have been performed between two organisms,

but very few have been performed on enough organisms to be able to build a phylogenetic tree.

The study of Sankoff *et al.* (1992) represents an example where a phylogenetic tree was constructed on the basis of gene order. Their comparison was performed using the mitochondrial genomes of 16 fungi and other eucaryotes. The mitochondrion is ideal to study genomic changes since its genome is small and consequently, contains an manageable number of genes. Sankoff *et al.* (1992) calculated the distances by determining the number of steps required to convert the gene order of a mitochondrial genome to another. The phylogenetic distances that were inferred from the 16 mitochondrial gene orders generally agreed with distances calculated from gene sequences.

My comparative analysis has shown that when comparing the qualitative tree based upon the similarities observed in gene organization to the actual distance tree, both trees are different. Nonetheless, the relationship between the *S. solfataricus* / *S. shibatae* subgroup was found identical in both trees. Therefore, conservation in gene organization between some species does seem to reflect their phylogenetic distance. My results overall, however, cast doubt on the value of gene order for phylogenetics, although I only studied a limited region of the genome.

At this point, answering if gene order can be used as a tool to construct trees is uncertain. Not enough studies have been performed. Hopefully, this will

change, once more sequences of complete genomes become available. A problem however that is becoming more obvious is the need for an effective way to calculate distances from gene orders. Sankoff *et al.* (1992) have developed software that will calculate the minimal number of rearrangements required to convert the order of one genome to another. Nonetheless, these calculations become unfeasable when pairs of genomes contain too many genes in common, especially when the organisms are not closely related. This will definitely become a problem when comparing gene orders of completely sequenced genomes that contain a couple of thousand genes. New computer programs will have to be developed to account for the complexity of these analyses. Alternatively, new methods to calculate the distances might have to be considered, such as performing statistics on neighboring pairs of genes.

IV. Evaluation of my novel comparative approach

IV.A. Variables that are influencing the findings

My comparative sequencing analysis constitutes the first of its kind. Using this novel approach, conservation in gene organization within the *Sulfolobales* was detected. My analysis focused on a selected region of the genome of seven *Sulfolobales*. When undertaking a comparison of this type, two important variables have to be considered: the reference gene chosen and the organisms chosen.

The region of the genome chosen will have an effect on the results obtained from the comparative analysis. Clearly, one has to be certain that the equivalent regions in the different organisms are actually being compared. In other words, orthologous genes, and not paralogous genes, have to be used as a reference in the genomes. For my study, I chose the 16S rRNA gene as my reference gene, since it is present as a single copy gene in the *Sulfolobales*. Once the reference gene is chosen, there are no guarantees that this gene will be within a region that is easily clonable, and at the same time, located in the middle of the cosmid clone for all organisms compared. For instance, I was unsuccessful at cloning the 16S rRNA gene region of *Sulfolobus hakonensis*, another organism that had been chosen for my comparison. Fortunately, the same problem did not occur with the other *Sulfolobales*.

The organisms selected for the comparative analysis, will also have an impact on the findings. In order to observe if rearrangements accumulate constantly over time, the organisms chosen for the analysis must have diverged at different times, and at the same time, must be related closely enough to observe some conservation in gene organization. Many comparative sequence analyses that have been performed previously, have shown that gene order was not conserved, mainly because the organisms chosen were too distantly related (Tatusov *et al.*, 1996; Fraser *et al.*, 1995; Mushegian and Koonin, 1996; Siefert *et al.*, 1997). In my study, I have chosen different species and strains in the order

Sulfolobales that are relatively closely related. Based on the 16S rRNA sequence and a clock rate of 1% substitutions per 50 million years, the different *Sulfolobales* chosen for my analysis were estimated to have diverged from each other, 25 to 550 million years ago (Ochman and Wilson, 1987). When initiating a project of this type, it is preferable to compare only two organisms at first, to verify the degree of conservation between these species. Then, one can decide if the other organisms in the comparison should be more or less closely related.

IV.B. Strengths and weaknesses of limiting a comparative analysis to a selected region of the genome

Concentrating my comparative analysis to a selected region of the genome carries advantages. A comparative sequencing analysis can be labor intensive, time consuming and expensive. By restricting my study to a 40-50 kbp region of the genome, the analysis becomes more realistic, and allows comparison of many organisms to each other in a fair amount of time. Since my main interest is to study gene organization, the region sequenced does not need to be error-free nor completely sequenced. Thus, by simply performing random sequencing, the cost of the experiment is lowered.

When performing a comparative analysis at the whole genome level, the analysis becomes very complex. A typical procaryotic genome contains approximately two thousand genes. Thus, comparing gene order of all these genes can be overwhelming, and impossible when the organisms compared are too distantly related. When concentrating our attention on one specific region of the

genome, the analysis is simplified, yet the high resolution at the sequence level is retained. Moreover, some comparative analyses previously performed have shown that although gene order was not conserved at the whole genome level, many short regions of conserved gene order could be observed (Tatusov *et al.*, 1996; Fraser *et al.*, 1995; Mushegian and Koonin, 1996; Siefert *et al.*, 1997). Thus, a comparison at a smaller scale might be more appropriate when comparing organisms that are not that closely related.

There are disadvantages of using only a section of the genome for a comparative analysis. Its biggest weakness is that this region might not necessarily be representative of the organization of the whole genome. Naturally, to resolve this problem, multiple comparative analyses of many sections of the genome could be performed. As a consequence however, the comparison would become a lot more involved and expensive. The gene arrangement observed in *S. tokodaii* could well be an example where the organization of this region does not represent that of the whole genome. Indeed, my analysis showed that *S. tokodaii* does not share any similarity in gene organization with the other *Sulfolobales* except for the rRNA operon. However, the phylogenetic distance based upon the 16S rRNA sequences of *S. tokodaii* and *S. acidocaldarius* showed that these species were closely related (Table 1). A possible scenario is that the many repeated sequences present in that particular region of *S. tokodaii*, were responsible for these rearrangements. What my comparative analysis fails to show, is whether the gene

arrangement in the region surrounding the 16S rRNA gene of the ancestor of the *Sulfolobales* was simply shifted further away at some stage in *S. tokodaii*, without altering the gene organization of that region. A possible means to determine if this region was actually moved in *S. tokodaii*, would be to hybridize the different genes present in the 7-8 kbp conserved block shared between the other *Sulfolobales*, against genomic digests of *S. tokodaii*. If the genes are linked to each other in one restriction fragment, then I could conclude that this region in *S. tokodaii* was not extensively shuffled around, but simply moved to another location of the genome.

Another important weakness of this type of comparative analysis, is that rearrangements in gene organization can be poorly quantified. In my study, a qualitative tree was built (Figure 9), simply on the basis of the similarities observed in gene organization in Figures 7 and 8. However, quantifying these distances based upon the similarities in gene order is not an easy task. In the course of my analysis, I did try to calculate similarity distances by dividing the average length of similarity in gene order shared between two species in kbp, by the total average length compared in kbp. However, this method turned out to be too imprecise, mainly because the contig maps are not the same length and because some sequences are missing between the different contigs. I then concluded that there was no acceptable way to quantify the distances based on the data obtained from this gene order comparison. If the similarities in gene organization could be quantified, then we could calculate the rate of

rearrangements in gene organization by plotting the similarities in gene order against the evolutionary distance based on the 16S rRNA sequences.

IV.C. An alternative comparative approach

A completely different approach could be taken, that would consider all the weaknesses of my comparative approach. Using this method, total genomic DNA of many closely related organisms would be nebulized and cloned into a plasmid vector. For each library constructed, many plasmids containing an insert of 2-3 kbp would be chosen. Then, the ends of the inserts would be used as probes against genomic digests of the different species, in order to detect if these two ends are linked in the various organisms. Compared to my analysis, this approach can quantify effectively the conservation in gene organization, it is representative of the whole genome, and more organisms can be compared since the labor and the cost are not as exigent.

V. New research directions projected from my findings

V.A. Comparative analyses of individual genes

Another benefit to my comparative analysis is being able to compare phylogenetically the different genes found in the conserved regions. The 7-8 kb conserved block of gene order observed in *S. solfataricus* P1/P2, *S. shibatae*, *S. acidocaldarius* and *M. sedula* can indeed be used for phylogenetic purposes. This conserved region is composed of 7 genes, whose sequences can be compared. By

using genes that are within a conserved region to build phylogenies, we can almost be certain that the genes being compared are orthologous. The 16S and 23S rRNA genes are the only genes that have been used so far to study the phylogeny of the *Sulfolobales* (Fuchs *et al.*, 1996; Trevisanato *et al.*, 1996). The relationship that would be obtained from the 7 phylogenies could therefore be compared to the 16S rRNA phylogeny. But in order to do so, the 7-8 kbp region in the different *Sulfolobales* needs to be completely sequenced and double-stranded. This additional part of my project is in fact already in progress.

When performing a comparative analysis of this type, many genes are being discovered and some may be of some interest to researchers. Indeed, some of the genes found in the region surrounding the 16S rRNA gene have already caught the attention of some researchers. For example, the type II DNA topoisomerase subunit A and B found in *S. acidocaldarius* and *M. sedula*, the ABC transporters found in *S. shibatae*, *S. acidocaldarius*, and *M. sedula*, and the coiled-coil protein found in *M. sedula* have raised some interests. Again if necessary, more sequencing will have to be performed to obtain the complete sequence of these genes.

V.B. The breakpoints

From my comparative analysis, breakpoints where rearrangements occurred can be studied. These breakpoints are important because they can give us an indication of the forces that are disrupting gene order. In order to identify these

rearrangement events, sequence signatures will have to be searched at the junctions where rearrangements occurred. The regions of breakpoints can easily be identified from Figure 8 and also from the dot plots (Appendix VI). For example, Figure 8 shows that *S. acidocaldarius* and *M. sedula* when compared to *S. solfataricus* P2, share a common breakpoint at one end of the 7-8 kbp conserved block region. The sequences from these three organisms in that particular region can then be aligned in order to investigate if any vestiges of the rearrangement can be found. There are two possibilities: either nothing will be found mainly because these breakpoints are located between genes, where the rate of substitutions is high or, sequence signatures will be found. In the latter case, their discovery could turn out to be very informative in understanding the forces that are maintaining and disrupting gene organization.

VI. Conclusion

Overall, my novel comparative approach was a success. I was able to detect conservation in gene organization in the region surrounding the 16S rRNA gene of *Sulfolobales* and construct a qualitative tree based on the different gene arrangements observed in that selected region of the genome. I concluded that rearrangements in that particular region of the genome followed a punctuated model of evolution and that the ancestor of the *Sulfolobales* had the *S. acidocaldarius* / *M. sedula* gene arrangement in the region surrounding the 16S

rRNA gene. Even though I was unable to draw firm conclusions regarding the use of gene order as a tool to build phylogenies, there is no doubt in my mind that gene order analyses will soon complement traditional sequences comparisons.

APPENDIX IA: Recipes of solutions

Denaturing solution:

0.5 M NaOH	20 g NaOH
1.5 M NaCl	87.6 g NaCl
	Bring volume to 1 liter with water.

Depurination solution:

0.25 M HCl	10.78 ml of concentrated HCl
	489.2 ml of water.

Hybridizing solution:

1M NaCl	29.2 g NaCl
50 mM Tris·Cl pH 7.6	25 ml 1 M Tris·Cl pH 7.6
5% w/v SDS	25 g SDS
50% w/v Formamide	250 ml Formamide
	Bring volume to 500 ml with water.

10X Klenow buffer:

0.5 M Tris·Cl pH 9.5
0.1 M MgCl2
0.05 M Dithiothreitol
1 mM Spermidine

10X Ligase buffer without ATP:

660 mM Tris·Cl
50 mM $MgCl_2$
10 mM Dithioerythritol

Neutralizing solution:

0.5 M Tris·Cl pH 7.4	500 ml of 1 M Tris·Cl pH 7.4
1.5 M NaCl	87.6 g NaCl
	Bring volume to 1 liter with water.

Phage dilution buffer: (per liter)

100 mM NaCl	5.8 g NaCl
17 mM $MgSO_4$	2 g $MgSO_4$
50 mM Tris·Cl pH 7.6	50 ml of 1 M Tris·Cl, pH 7.6
0.01% gelatin	5 ml of 2% gelatin
	Bring volume to 1 liter with distilled water.
	Autoclave.

10X RE:

500 mM Tris·Cl, pH 8
500 mM KCl
100 mM $MgCl_2$
100 mM Dithiothreitol

10X RP-C:

200 mM Tris·Cl, pH 7.6
100 mM $MgCl_2$
50 mM Dithiothreitol
660 µM each dG, dA and dT
25% glycerol

Size marker (λ - BstE II - Xho I -Xba I):

λ DNA digested with BstE II:
120 µg of λ DNA
80 µl of 10X RE buffer
20 µl of 4 M NaCl
10 µl BstE II (100 U)
Bring volume to 800 µl with water.
Incubate at 65 °C for 3 hours.

λ DNA digested with Xho I:
40 µg of λ DNA
80 µl of 10X RE buffer
5 µl Xho I (50 U)
Bring volume to 800 µl with water.
Incubate at 37°C for 3 hours.

λ DNA digested with Xba I
40 µg of λ DNA
80 µl of 10X RE buffer
8 µl of Xba I (80 U)
Bring volume to 800 µl with water.
Incubate at 37°C for 3 hours.

Mix together in a tube:
100 µl of λ-BstE II
100 µl of λ-Xho I
100 µl of λ-Xba I
100 µl of TE
100 µl of 0.25 M EDTA
150 µl of loading buffer

20X SSC:

3 M NaCl
0.3 M Sodium citrate

350.6 g NaCl
176.4 g Sodium citrate
Bring volume to 2 liters with water.

STE:

1% N-lauryl sarcosine	0.2 g N-lauryl sarcosine
50 mM Tris·Cl pH 8	1 ml of Tris·Cl pH 8
50 mM EDTA pH 8	2 ml of 0.5 M EDTA pH 8
	17 ml of water

40X TAE:

2M Tris	484.5 g Tris
80 mM EDTA	59.9 g EDTA
0.8 M Sodium acetate	217.8 g Sodium acetate
	Add 1857.6 ml of water.
	Adjust pH to 7.2 by adding 142.4 ml of acetic acid.

TE:

10 mM Tris·Cl pH 8	10 ml of 1M Tris·Cl pH 8
1 mM EDTA pH 8	2 ml of 500 mM EDTA
	Bring volume to 1 liter with water.
	Autoclave.

TEG:

20 mM Tris·Cl pH 8	1 ml of 1 M Tris·Cl pH 8
50 mM EDTA	5 ml of 0.5 M EDTA
1% glucose	5 ml of 10% glucose
	Bring volume to 50 ml with water.
	Filter sterilize.

Transfer solution:

0.4 M NaOH	16 g of NaOH
	Bring volume to 1 liter with water.

Prewash solution:

5X SSC	500 ml of 20X SSC
0.5% SDS	10 g of SDS
1mM EDTA	0.74 g of EDTA
	Bring volume to 2 liter with water.

APPENDIX IB: Recipes of Media

***Sulfolobales* media:**

1X broth: (per 50 ml)

500 µl of 100X salts solution
250 µl of 200X salts solution
50 µl of 1000X salts solution
500 µl 20% sucrose
500 µl 10% yeast extract
100-250 µl 1:50 H_2SO_4:H_2O

Solid media: (per liter)

10 ml 100X salts
5 ml 200X salts
1 ml 1000X salts
10 ml 20% sucrose
10 ml 10% yeast extract
6 ml 0.5 M $CaCl_2$
10 ml 1 M MgCl2
455 ml H_2O
Heat the salt medium to 60°C
Add 8 g gelrite already dissolved in 500 ml of water by heating
Add 1:50 H_2SO_4:H_2O to bring the pH to 3.5-4.
Pour plates.

100X salts solution: (per 2 liter)

260 g $(NH_4)_2SO_4$
50 g $MgSO_4 \cdot 7H_2O$
Autoclave.
Add 4 g $FeCl_3 \cdot 6H_2O$ filter sterilized
Add 3 ml 50% H_2SO_4

200X salts solution: (per liter)

56 g KH_2PO_4
36 ml 10 mg/ml $MnCl_2 \cdot 4H_2O$
4.4 ml 10 mg/ml $ZnSO_4 \cdot 7H_2O$
1 ml 10 mg/ml $CuCl_2 \cdot 2H_2O$
0.6 ml 10 mg/ml $VOSO_4 \cdot 2H_2O$
0.2 ml 10 mg/ml $CoSO_4$
90 ml 10 mg/ml $Na_2B_4O_7 \cdot 10H_2O$
0.6 ml 10 mg/ml $Na_2MoO_4 \cdot 2H_2O$
Autoclave.
Add 5 ml 50% H_2SO_4

1000X salts solution: (per 200 ml)
14 g $CaCl_2 \cdot 2H_2O$

SOB medium: (per 400 ml)
8 g tryptone
2 g yeast extract
0.2 g NaCl
Autoclave
Add 4 ml of sterile 1 M $MgCl_2$
Add 4 ml of sterile 1 M $MgSO_4$

SOC medium: (per 40 ml)
39 ml of SOB medium
0.8 ml of filter sterilized 20% w/v glucose
Filter sterilize just before use.

YT-Ampicillin broth: (per liter)
YT broth
Autoclave
Cool to 55°C
Add 70 mg of filter-sterilized ampicillin.

YT-Ampicillin-Xgal-IPTG plates: (per 500 ml)
YT agar
Autoclave
Cool to 55°C
Add 35 mg of filter-sterilized ampicillin.
Add 2 ml of 2% Xgal
Add 0.4 ml of filter-sterilized 1 M IPTG
Pour into petri dishes.

YT broth: (per liter)
8 g tryptone
5 g yeast extract
5 g NaCl
Autoclave.

YT-Kanamycin broth: (per liter)
YT broth
Autoclave
Cool to 55°C
Add 30 mg of filter-sterilized kanamycin.

YT-Kanamycin broth or plates: (per 500 ml)

YT agar
Autoclave
Cool to 55°C
Add 15 mg of filter-sterilized kanamycin.
Pour into petri dishes.

YT plates: (per liter)

YT broth
15 g agar
Autoclave.

APPENDIX II: Table of contig names (C), length of contigs (L), orientation (O) and position of contigs on the maps for *S. acidocaldarius*, *S. shibatae*, *S. solfataricus* MT4, *S. tokodaii* and *M. sedula*.

S. acidocaldarius				*S. shibatae*				*S. solfataricus* MT4				*S. tokodaii*				*M. sedula*			
Map position	C	L	O	Map position	C	L	O	Map position	C	L	O	Map position	C	L	O	Map position	C	L	O
1-506	91	506	+	1-1980	113	1980	+	1-1334	202	1334	+	1-225	13	225	+	1-533	228	533	+
598-977	19	379	-	2133-6750	126	4617	-	1431-2782	165	1351	-	830-1421	2	591	+	561-1952	143	1391	+
1610-3212	29	1602	-	6857-7375	154	518	-	3137-3630	45	493	-	1831-4927	29	3096	+	2124-2551	105	427	-
3290-3900	112	610	-	7412-7915	2	503	+	3875-4304	9	429	+	5012-5459	158	447	+	2651-4126	294	1475	-
3960-4363	151	403	-	8000-13101	102	5101	+	4334-4971	5	637	+	5610-6983	55	1373	+	4387-5128	206	741	+
4438-4924	191	486	+	13195-14040	129	845	+	5157-5556	19	399	+	7062-7546	153	484	-	5171-7006	320	1835	-
5155-6488	69	1333	+	14054-14888	200	834	-	5606-6382	20	776	-	7630-8779	176	1149	+	10006-21699	47	11693	-
6513-8793	110	2280	-	15326-15785	85	459	+	6389-7094	53	705	+	8809-10557	103	1748	-	21717-22135	84	418	+
8890-10208	111	1318	-	16500-17244	116	744	+	7139-8549	57	1410	+	11480-12103	102	623	-	22145-23381	232	1236	-
10277-10770	88	493	-	17662-19148	14	1486	+	8571-10557	64	1986	-	12153-12880	200	727	-	23382-26182	192	2800	+
11077-12471	43	1394	+	19230-21398	162	2168	+	10721-14412	18	3691	-	13307-13769	11	462	+	26299-38131	265	11832	-
12617-14024	107	1407	-	21450-25903	177	4453	+	14516-15557	91	1041	-	13887-16027	219	2140	+	41131-41822	21	691	+
14129-15461	18	1332	-	26002-28394	61	2392	+	15626-15881	76	255	-	16206-17056	73	850	+	41988-46265	114	4277	-
15643-17491	78	1848	+	28500-29419	21	919	+	16883-17621	3	738	+	17401-18858	39	1457	-	46715-48297	255	1582	-
17573-18467	146	894	-	30010-30452	57	442	-	18635-19375	67	740	-	18928-28195	109	9267	-	48302-49134	70	832	+
18655-19174	198	519	-	30500-30957	140	457	-	19716-20189	103	473	+	28315-31416	175	3101	-	50702-51914	69	1212	+
19195-20577	182	1382	-	31813-32552	118	739	+	20197-22184	17	1987	+	31478-32530	58	1052	-	52369-54829	38	2460	+
22088-22695	108	607	-	32636-34151	207	1515	+	22357-23596	87	1239	+	32642-33084	91	442	-	54837-55398	268	561	-
23556-24009	27	453	+	34383-35450	28	1067	-	23638-27007	101	3369	-	34194-34633	76	439	-	55459-59259	183	3800	+
24283-24702	147	419	+	35855-36273	32	418	-	27741-28053	182	312	-	34952-38504	59	3552	+				
25283-28812	160	3529	+	36645-37644	134	999	+	30621-31493	170	872	+	38544-39631	119	1087	-				
29008-31218	116	2210	-	37695-41110	69	3415	-	31501-31928	195	427	+	39652-40379	214	727	+				
31390-37477	71	6087	+	41138-41822	186	684	+	32411-35630	179	3219	+	40574-41474	9	900	-				
				41900-42713	204	813	+	35660-39485	98	3825	+	41643-42474	46	831	-				
				43016-44062	13	1046	-	39488-39943	106	455	+	42580-43333	146	753	+				
								40047-41616	183	1569	-	43485-43945	95	460	-				
								41987-43182	155	1195	+	44314-45175	208	861	-				
								43241-44731	185	1490	+	46230-46743	229	513	-				
								44772-47992	188	3220	-								
								48098-48580	157	482	-								

APPENDIX IIIA: Results obtained from BLASTN and BLASTX searches performed on the contig sequences of *S. solfataricus* MT4.

Contig name	Position of contig on map	BLASTN searches against NCBI database	BLASTX searches against NCBI database
mt202	1-1334	196-497, 1112-1229: *S. solfataricus* 100 Kb fragment (e-43)	
mt165	1431-2782		
mt45	3137-3630		
mt9	3875-4304		3886-4098: Neuronal glycoprotein / mouse (e-5)
mt5	4334-4971		4368-4955: Orf C4L / *Swinepox virus* (e-8) [also matches beta scruin]
mt19	5157-5556		5176-5523: Sulfate transport system protein / *Chlorella vulgaris* (e-13)
mt20	5606-6382		6012-6365: MJ 1186 / *M. jannachii* (e-8)
mt53	6389-7094		
mt57	7139-8549	7337-7676: *Anabaena* sp. polyketide synthase and ketoacyl reductase genes (e-5)	7192-7827: Hypothetical oxidoreductase in GLNQ-ANSR intergenic region Yqjq / *Bacillus subtilis* (e-32) 7883-8548: Formate DH alpha subunit / *M. jannachii* (e-23)
mt64	8571-10557	8962-9243: *M. jannachii* section 117 of genome (e-18) 9533-9862: *M. jannachii* section 1 of genome (e-8)	8673-9791: FdhA / *Methanobacterium thermoformicicum* (e-68) [also matches formate DH α subunit]
mt18	10721-14412	11869-13397: *S. solfataricus* 100 Kb DNA fragment (≈0)	12096-13115: Orf c02014 / *S. solfataricus* (e-158) 13741-14409: Formate hydrogenlyase subunit 4 / *E. coli* (e-19)
mt91	14516-15557		
mt76	15626-15881		15631-15798: Z74410 unknown / *M. tuberculosis* (e-13) [also matches NADH DH]
mt3	16883-17621		16883-17538: Formate hydrogenlyase subunit 5 / *M. jannachii* (e-7)
mt67	18635-19375	19116-19366: *S. solfataricus* insertion element ISC 1217 (e-71)	18690-19109: HyfB / *E. coli* (e-20) [also matches formate hyd s3 & NADH ubiquinone oxidored. Chain 4] 19126-19366: ISC 1217 / *S. solfataricus* (e-38)
mt103	19716-20189	19716-20189: *S. solfataricus* insertion element ISC 1217 (e-176)	19716-20189: ISC 1217 / *S. solfataricus* (e-93)
mt17	20197-22184		21489-22157: Thiosulfate sulfurtransferase / *Mycobacterium leprae* (e-56)
mt87	22357-23596		23503-23592: Nickel transport protein / *Alcaligenes eutrophus* (e-7)
mt101	23638-27007	24541-25691: *S. solfataricus* insertion element ISC 1217 (≈0)	23638-24363: Nickel transport protein / *Alcaligenes eutrophus* (e-57) 24617-25678: ISC 1217 / *S. solfataricus* (e-242)
mt182	27741-28053	27741-28053: *S. solfataricus* 23S rRNA gene (e-112)	
mt170	30621-31493	30621-31493: *S. solfataricus* 16S rRNA gene (≈0)	
mt195	31501-31928	31501-31897: *S. solfataricus* 16S rRNA gene (e-148)	
mt179	32411-35630	32959-33349, 33445-35490: *S. solfataricus* 100 Kb DNA fragment (e-26) 33350-33444: *S. solfataricus* 16 rRNA gene (e-18)	
mt98	35660-39485		35869-36441: Orf c01015 / *S. solfataricus* (e-8)
mt106	39488-39943		
mt183	40047-41616		
mt155	41987-43182		
mt185	43241-44731		
mt188	44772-47992		46359-47348: Acetylornithine deacetylase / *E. coli* (e-18)
mt157	48098-48580		

APPENDIX IIIB: Results obtained from BLASTN and BLASTX searches performed on the contig sequences of *S. shibatae*.

Contig name	Position of contig on map	BLASTN searches against NCBI database	BLASTX searches against NCBI database
shi113	1-1980		1004-1672: Proliferating cell nuclear antigen homologue / *M. jannachii* (e-18)
shi126	2133-6750	6663-6745: *S. shibatae* tRNA nucleotidyltransferase (e-12)	2269-2508: C34 E10.2 gene product / *C. elegans* (e-14) 2572-3417: Ribose phosphate pyrophosphokinase / *M. jannachii* (e-44) 3945-4420: Hyp. protein in NIFH2 5' region / *Methanococcus thermolithotrophicus* (e-20) [also matches MJ0710] 4379-5167: MJ1502 / *M. jannachii* (e-38) 5402-5692: MJ1564 / *M. jannachii* (e-7) 5694-6227: MJ1399 / *M. jannachii* (e-25) 6273-6750: MJ1568 / *M. jannachii* (e-20)
shi154	6857-7375	6857-7375: *S. shibatae* tRNA nucleotidyltransferase (e-208)	6857-7375: tRNA nucleotidyltransferase / *S. shibatae* (e-116)
shi2	7412-7915	7412-7828: *S. shibatae* tRNA nucleotidyltransferase (e-160)	7415-7771: tRNA nucleotidyltransferase / *S. shibatae* (e-69)
shi102	8000-13101	8809-9146: *Sulfolobus* sp. B12 tRNA gene with integrated SSV1 genome (e-66) 11108-11400: *M. jannachii* section 118 of genome (e-5)	8808-9041: Probable integrase *Sulfolobus* particle SSV1 (e-26) 11137-11628: ABC transporter / *Synechocystis* sp. (e-15) 12720-12977: Titin / rabbit (e-11)
shi129	13195-14040		13334-13837: Beta scruin / *Limulus polyphemus* (e-6)
shi200	14054-14888		14209-14563: Sulfate transport system permease / *Synechocystis* sp. (e-11)
shi85	15326-15785		
shi116	16500-17244		16588-17202: Formate DH alpha subunit / *M. jannachii* (e-24)
shi14	17662-19148	17871-18003: *Haemophilus influenzae* section 1 of genome (e-7)	17748-18509: Formate DH alpha subunit / *Wolinella succinogenes* (e-53) 18600-19106: Hypothetical protein / *Acidianus ambivalens* (e-7) [MJ1582]
shi162	19230-21398		20441-21291: Formate hydrogenlyase subunit 4 / *E. coli* (e-24)
shi177	21450-25903		22232-23282: Z74410 unknown / *M. tuberculosis* (e-49) [also matches NADH ubiquinone oxidoreductase chain 4] 24703-25062: Formate hydrogenlyase subunit 7 / *M. jannachii* (e-28) [also matches NADH ubiquinone oxidoreductase subunit K] 25559-25891: Formate hydrogenlyase subunit 5 / *M. jannachii* (e-9)
shi61	26002-28394		27273-28034: Thiosulfate sulfurtransferase / *Mycobacterium leprae* (e-56)
shi21	28500-29419		28604-29269: Nickel transport protein / *Helicobacter pylori* (e-66)
shi57	30010-30452	30010-30452: *S. shibatae* 23S rRNA gene (e-170)	
shi140	30500-30957	30500-30957: *S. shibatae* 23S rRNA gene (e-166)	
shi118	31813-32552	31813-32552: *S. shibatae* 23S rRNA gene (e-287)	
shi207	32636-34151	32662-34151: *S. shibatae* 16S rRNA gene (≈0)	
shi28	34383-35450	34383-35450: *Sulfolobus* sp. strain B12 16S rRNA gene upstream region (≈0)	
shi32	35855-36273	35855-35986: *Sulfolobus* sp. strain B12 16S rRNA gene upstream region (e-21)	
shi134	36645-37644		
shi69	37695-41110		38808-39392: Hypothetical protein / *Synechocystis* sp. (e-24) [also matches orf c01038 / *S. solfataricus*] 39919-40460: ATP binding protein / *Streptococcus pneumoniae* (e-28)
shi186	41138-41822		
shi204	41900-42713		42154-42630: Putative maltose binding protein / *Streptomyces coelicolor* (e-5)
shi13	43016-44062		43329-44062: Quinolate synthetase / *M. jannachii* (e-52)

APPENDIX IIIC: Results obtained from BLASTN and BLASTX searches performed on the contig sequences of *S. acidocaldarius*.

Contig name	Position of contig on map	BLASTN searches against NCBI database	BLASTX searches against NCBI database
a91	1-506		
a19	598-977		
a29	1610-3212		2216-2917: Translation initiation factor, eIF-2 subunit alpha / *M. jannachii* (e-60) 2922-3098: 30S ribosomal protein S27E / *M. jannachii* (e-17) [also matches 40S ribosomal protein]
a112	3290-3900		
a151	3960-4363		3960-4051: MJ0839 / *M. jannachii* (e-7)
a191	4438-4924		
a69	5155-6488		5676-6188: C34 E10.2 gene product / *C. elegans* (e-28)
a110	6513-8793		6781-7090: Ribose-phosphate pyrophosphokinase / *M. jannachii* (e-28) 7600-8016: Hypothetical protein in NIFH2 5' region / *M. thermolithotrophicus* (e-16) 8010-8793: MJ1502 / *M. jannachii* (e-48)
a111	8890-10208		9028-9312: MJ1564 / *M. jannachii* (e-9) 9324-9839: MJ1399 / *M. jannachii* (e-31)
a88	10277-10770	10372-10526: *M. jannachii* section 139 of genome (e-7) 10564-10758: *S. shibatae* tRNA nucleotidyltransferase (e-20)	10354-10539: MJ1568 / *M. jannachii* (e-13) 10564-10758: tRNA nucleotidyltransferase / *S. shibatae* (e-24)
a43	11077-12471		11077-11842: tRNA nucleotidyltransferase / *S. shibatae* (e-91)
a107	12617-14024	12895-12997: *S. solfataricus* 56 Kb fragment (e-7)	13855-14023: Titin / human (e-6)
a18	14129-15461		14205-15026: Beta scruin / *Limulus polyphemus* (e-10)
a78	15643-17491	15659-16074: *C. elegans* cosmid C06B3 (e-6)	15649-16069: Hypothetical oxidoreductase in GLNQ-ANSR intergenic region Yqjq / *B. subtilis* (e-18) 16120-16617: MJ1582 / *M. jannachii* (e-17)
a146	17573-18467	17582-17692: *L. lactis* genes for ATP-binding protein (e-6)	17585-18016: ATP binding protein / *B. subtilis* (e-29)
a198	18655-19174	18660-19174: *S. acidocaldarius* 16S and 23S rRNA genes (e-198)	
a182	19195-20577	19195-20577: *S. acidocaldarius* 16S and 23S rRNA genes (≈0)	
a108	22088-22695	22088-22695: *S. acidocaldarius* 23S rRNA gene (e-244)	
a27	23556-24009	23556-24009: *S. acidocaldarius* 16S rRNA gene (e-182)	
a147	24283-24702	24283-24697: *S. acidocaldarius* 16S rRNA gene (e-138)	
a160	25283-28812	25283-25716: *S. acidocaldarius* 16S and 23S rRNA genes (e-158)	27423-28722: 2-isopropylmalate synthase / *M. jannachii* (e-91)
a116	29008-31218		29214-29771: Uridylate kinase / *M. jannachii* (e-44) 30161-30993: 7-alpha hydroxysteroid DH / *Clostridium sordellii* (e-19)
a71	31390-37477	31848-33439: *S. acidocaldarius* ffh gene (≈0)	31734-31988: Orf-2 / *Acidianus ambivalens* (e-29) 31993-33330: 54 homologue of SRP54 / *S. acidocaldarius* (e-301) 33329-33733: Translation initiation factor aIF-5A / *S. acidocaldarius* (e-86) 34070-35221: Type II DNA topoisomerase subunit A / *S. shibatae* (e-217) 35339-36819: Type II DNA topoisomerase subunit B / *S. shibatae* (e-259) 36867-37322: Deoxypusine synthase / yeast (e-34)

APPENDIX IIID: Results obtained from BLASTN and BLASTX searches performed on the contig sequences of *S. tokodaii*.

Contig name	Position of contig on map	BLASTN searches against NCBI database	BLASTX searches against NCBI database
t13	1-225		
t2	830-1421		
t29	1831-4927		3956-4697: 5-methycytosine specific restriction enzyme B / *E. coli* (e-9)
t158	5012-5459		
t55	5610-6983		6584-6982: Transposase / *Vibrio cholerae* (e-26) [IS 200]
t153	7062-7546		
t176	7630-8779		7631-7688, 7899-8057: Orf ID:o271#4 / *E. coli* (e-7)
t103	8809-10557		
t102	11480-12103		
t200	12153-12880		
t11	13307-13769		
t219	13887-16027	15007-15976: *S. solfataricus* 100 Kb fragment (e-125)	13989-14936: Orf c06026 / *S. solfataricus* (e-24) (ISC 1228) 15155-16027: Orf c06024 / *S. solfataricus* (e-155)
t73	16206-17056	16386-17056: *S. solfataricus* 100 Kb fragment (e-107)	16207-17056: Orf c06024 / *S. solfataricus* (e-139)
t39	17401-18858	17436-17930: *S. solfataricus* 100 Kb fragment (e-90)	17436-17930: Orf c06024 / *S. solfataricus* (e-87)
t109	18928-28195		19412-20549: Membrane transport protein / *B. subtilis* (e-39) 23297-23426, 20935-21153: Ferredoxin / *S. acidocaldarius* (e-44) 24579-25302: IS 1000 / *Thermus aquaticus* (e-11)
t175	28315-31416		
t58	31478-32530	32238-32479: *S. solfataricus* 23S rRNA gene (e-47)	
t91	32642-33084	32642-33084: *Stygiolobus azoricus* 23S rRNA gene (e-139)	
t76	34194-34633	34244-34633: *Stygiolobus azoricus* 23S rRNA gene (e-130)	
t59	34952-38504	35166-37130: *S. acidocaldarius* 16S rRNA gene (≈0)	
t119	38544-39631		
t214	39652-40379		
t9	40574-41474		41061-41444: Acetyl-coenzyme A synthetase / *B. subtilis* (e-8)
t46	41643-42474		41645-42460: Acetyl-coenzyme A synthetase / *B. subtilis* (e-60)
t146	42580-43333		
t95	43485-43945		
t208	44314-45175		44396-44666: Lipoate protein ligase lplA / *Mycoplasma pneumoniae* (e-5)
t229	46230-46743		

APPENDIX IIIE: Results obtained from BLASTN and BLASTX searches performed on the contig sequences of *M. sedula.*

Contig name	Position of contig on map	BLASTN searches against NCBI database	BLASTX searches against NCBI database
m228	1-533		
m143	561-1952		701-1564: F56H9.1 / *C. elegans* (e-7)
m105	2124-2551		
m294	2651-4126		3627-3791: Oligosaccharyl transferase STT3 subunit homolog / *C. elegans* (e-10)
m206	4387-5128		
m320	5171-7006		5783-6043: 60S ribosomal protein L44 / *Chlamydomonas reinhardtii* (e-17) 6076-6252: 30S ribosomal protein S27E / *M. jannachii* (e-14) 6257-6943: Translation initiation factor, eIF-2, subunit alpha / *M. jannachii* (e-57)
m47	10006-21699	20215-21129: *S. shibatae* tRNA nucleotidyltransferase (e-77)	14413-15077: Proliferating cell nuclear antigen / *M. jannachii* (e-9) 15388-15900: C34E10.2 gene product / *C. elegans* (e-22) 15961-16770: Ribose-phosphate pyrophosphokinase / *M. jannachii* (e-35) 17316-17741: Hypothetical protein in NIFH2 5' region / *M. thermolithotrophicus* (e-9) 17799-18581: MJ 1502 / *M. jannachii* (e-42) 18661-19080: MJ 1564 / *M. jannachii* (e-9) 19128-19598: MJ 1399 / *M. jannachii* (e-23) 19737-20173: MJ 1568 / *M. jannachii* (e-21) 20216-21445: tRNA nucleotidyltransferase / *S. shibatae* (e-145)
m84	21717-22135		21942-22028: Hypothetical oxidoreductase in CYSP 5' region / *S. typhimurium* (e-4)
m232	22145-23381	22723-22841: *S. solfataricus* 56 Kb DNA fragment (e-7)	23111-23236: Sulfate transport system permease / *Synechococcus* sp.(e-3)
m192	23382-26182		23774-24415: MJ1186 / *M. jannachii* (e-23) 25566-26099: Hypothetical oxidoreductase in RTP 5' region (ORF238) / *B. subtilis* (e-33) [Hyp. oxidoreductase in GI.NQ-ANSR intergenic region]
m265	26299-38131	37709-38131: *Acidianus brierleyi* 23S rRNA gene (e-114)	26768-27403: Possible 2-hydroxyhepta-2,4- diene-1,7- dioate isomerase / *Mycobacterium leprae* (e-13) 27660-28953: 4-hydroxyphenylacetate-3 hydrolase / *Klebsiella pneumoniae* (e-38) 29024-29965: Homoprotocatechuate 2,3- dioxygenase / *Brevibacterium fuscum* (e-57) 30093-31418: Succinate-semialdehyde DH / *Rhizobium* sp. (e-91) 33063-33593: Membrane transporter (e-18) 36809-37441: Myosin heavy chain (e-14)
m21	41131-41822	41131-41822: *M. sedula* 16S rRNA gene (e-274)	
m114	41988-46265	41988-43422: *M. sedula* 16S rRNA gene (≈0)	45540-45875: Orf c02003 / *S. solfataricus* (e-18)
m255	46715-48297	47370-47520: *Thermus aquaticus* thermophilus isopropylmalate synthase (e-13)	46885-48097: 2-isopropylmalate synthase (MJ1195) / *M. jannachii* (e-104)
m70	48302-49134		48530-49063: Uridylate kinase homolog / *M. jannachii* (e-42)
m69	50702-51914		50721-51026: 7-alpha hydroxysteroid DH / *E. coli* (e-10)
m38	52369-54829	53822-54825: *A. ambivalens* flh gene (e-160)	52381-52969: 3-hydroxyisobutyrate DH / rat (e-17) 53769-54035: Orf-2 / *A. ambivalens* (e-31) 54031-54825: 54 homologue of SRP54 / *A. ambivalens* (e-146)
m268	54837-55398	54842-55398: *A. ambivalens* flh gene (e-94)	54850-55371: 54 homologue of SRP54 / *A. ambivalens* (e-85)
m183	55459-59259	55459-55765: *S. acidocaldarius* gene for hypusine-containing protein (e-42) 56066-58880: *S. shibatae* top6B and top6A genes (≈0)	55461-55770: Translation initiation factor aIF-5A / *S. acidocaldarius* (e-45) 56135-57444: Type II DNA topoisomerase subunit A / *S. shibatae* (e-220) 57337-58881: Type II DNA topoisomerase subunit B / *S. shibatae* (e-261) 58937-59227: Deoxyhypusine synthase / *Homo sapiens* (e-28)

APPENDIX IVA: Results obtained from LOCALFASTA and TBLASTX searches performed on the contig sequences of *S. solfataricus* MT4 against sequences from *S. solfataricus* P2.

Contig name	Position of contig on map	LOCAL FASTA searches against *S. solfataricus* P2 database	TBLASTX searches against *S. solfataricus* P2 database
mt202	1-1334	233-481: c15 [13287-13037] (800) also matches: c51 (783), c38 (783), c21 (620), c6 (572), c24 (471)	243-491: c15 [13277-13028] (e-52) also matches: c51 (e-35), c38 (e-35), c13 (e-31), c21 (e-26), c6 (e-24), c24 (e-18), c1 (e-18)
mt165	1431-2782	1877-2781: c24 [27916-28805] (1014) also matches: c99 (679), c42 (549), c18 (549)	2058-2782: c24 [28083-28806] (e-45) also matches: c42 (e-34), c18 (e-33), c99 (e-16)
mt45	3137-3630	3137-3630: c23 [36468-35976] (1468)	3137-3630: c23 [36468-35976] (e-94)
mt9	3875-4304	3875-4301: c23 [35744-35321] (1203)	3875-4302: c23 [35744-35320] (e-79)
mt5	4334-4971	4335-4970: c23 [35286-34652] (1777)	4334-4971: c23 [35287-34651] (e-136)
mt19	5157-5556	5157-5556: c23 [34481-34085] (1193)	5157-5549: c23 [34481-34090] (e-71)
mt20	5606-6382	5607-6381: c23 [34069-33295] (2306)	5606-6381: c23 [34069-33296] (e-135)
mt53	6389-7094	6389-7093: c23 [33219-32513] (1979)	6389-7037: c23 [33220-32567] (e-120)
mt57	7139-8549	7139-8549: c23 [32466-31055] (3728) 7156-8549: c19 [1-1396] (3685)	7139-8549: c23 [32466-31055] (e-280) 7156-8549: c19 [1-1395] (e-276) also matches: c34 (e-20), c18 (e-10), c52 (e-10)
mt64	8571-10557	8572-10557: c23 [30982-29001] (4924) 8572-10557: c19 [1469-1986] (4924) also matches: c18 (772)	8587-10534: c23 [30968-29022] (≈0) 8587-10534: c19 [1483-3429] (≈0) also matches c18 (e-62)
mt18	10721-14412	11873-13393: c42 [22791-21273] (6054) also matches: c2 (2374), c34 (2368), c22 (2366), c40 (2361), c1 (2357), c18 (2357), c52 (2356), c24 (2335), ISC 1439 orf (1774)... 13542-14412: c23 [27967-27102] (2395) 13606-14412: c19 [4540-5349] (2395)	11868-13319: c42 [22796-21345] (≈0) also matches: c99 (e-204), c40 (e-204), c2 (e-203), c18 (e-202), c31 (e-201), c22 (e-201), c34 (e-201), c1 (e-200)... 12093-13115: ISC 1439orf [966-22] (e-159) 10721-10974, 13573-14412: c23 [28761-28508, 27938-27102] (e-182) 10721-10974, 13573-14412: c19 [3690-3943, 4513-5349] (e-183)
mt91	14516-15557	14516-15557: c23 [26965-25927] (2627) 14516-15557: c19 [5486-6524] (2627)	14546-15556: c23 [26936-25928] (e-146) 14546-15556: c19 [5515-6523] (e-146)
mt76	15626-15881	15630-15881: c23 [25876-25621] (664) 15626-15881: c19 [6570-6830] (664)	15628-15881: c23 [25879-25623] (e-36) 15628-15881: c19 [6572-6830] (e-36)
mt3	16883-17621	16883-17621: c23 [24464-23735] (1834) 16883-17621: c19 [7987-8716] (1834)	16883-17620: c23 [24464-23736] (e-126) 16883-17620: c19 [7987-8715] (e-126)
mt67	18635-19375	18636-19115: c23 [22690-22211] (1458) 18641-19116: c19 [9765-10240] (1458) 19116-19375: c17 [39734-39475] (879) also matches: c24 (879), c1 (870), c31 (852), c15 (852), ISC 1217orf (824)	18635-19115: c23 [22691-22211] (e-96) 18635-19115: c19 [9760-10239] (e-96) 19116-19375: c17 [39734-39477] (e-41) 19123-19375: ISC 1217 orf [1065-815] (e-40)
mt103	19716-20189	19716-20189: c24 [28914-29385] (1876) also matches: c1 (1802), c31 (1788), c6 (1788), c13 (e-1760), c17 (1737), ISC 1217 (1690), c15 (610)	19716-20188: c24 [28914-29384] (e-105) 19745-20189: ISC 1217 orf [1-444] (e-99) also matches: c1 (e-98), c6 (e-97), c31 (e-97), c13 (e-96), c17 (e-54)
mt17	20197-22184	20234-22184: c23 [22217-20276] (5462) 20235-22184: c19 [10234-12175] (5462)	20231-22182: c23 [22223-20278] (≈0) 20231-22182: c19 [10228-12173] (≈0)
mt87	22357-23596	23273-23591: c23 [20031-19714] (886) 23273-23591: c19 [12420-12737] (886)	23268-23592: c23 [20037-19713] (e-48) 23268-23592: c19 [12414-12738] (e-48) 22597-22783: c97 [1620-1431] (e-7)

mt101	23638-27007	24540-25688: c6 [5116-3968] (4566) also matches: c1 (4572), c31 (4550), c24 (4358), c17 (4324), c13 (4338), ISC 1217 orf (4211), c15 (938) 23638-24365: c23 [19652-18924] (2020) 23638-24367: c19 [12799-13527] (2020)	24540-25689: c6 [5116-3967] (e-272) 24618-25679: ISC 1217 orf [2-1063] (e-253) also matches: c1 (e-271), c31 (e-268), c13 (e-253), c24 (e-251), c17 (e-15)... 23648-24366, 26916-27007: c23 [19652-18924, 18589-18499] (e-142) 23638-24366, 26916-27007: c19 [12799-13527, 13862-13952] (e-142)
mt182	27741-28053	27741-28053: c23 [17577-17266] (1177) 27741-28053: c19 [14874-15185] (1777)	27741-28053: c23 [17576-17266] (e-67) 27741-28053: c19 [14874-15185] (e-67)
mt170	30621-31493	30621-31493: c23 [14960-14091] (3273) 30621-31493: c19 [17491-18360] (3273)	30623-31493: c23 [14958-14091] (e-193) 30623-31492: c19 [17493-18360] (e-193)
mt195	31501-31928	31501-31928: c23 [14082-13658] (1553) 31501-31928: c19 [18369-18793] (1553)	31502-31913: c23 [14081-13671] (e-86) 31502-31913: c19 [18370-18780] (e-86)
mt179	32411-35630	33217-33761: c97 [6929-7462] (468) also matches: c49 (468), c38 (430), c51 (430)	33848-35221: c24 [16635-15296] (e-94) also matches: c13 (e-21), c97 (e-16), c22 (e-15), c49 (e-15), c27 (e-13), c99 (e-12), c14 (e-12)
mt98	35660-39485	36626-37086: c13 [5004-4528] (1644) also matches: c34 (1649), c15 (1652), c25 (1662), ISC 1058orf (1882)	36441-37493: c13 [5271-4106] (e-172) 36569-37430: ISC 1058 orf [900-1] (e-160) also matches: c15 (e-168), c25 (e-168), c34 (e-149)
mt106	39488-39943		
mt183	40047-41616		
mt155	41987-43182		
mt185	43241-44731	43870-44731: c23 [8944-8098] (2593) 43848-44731: c19 [23485-24353] (2593) 44326-44699: c40 [18172-18545] (947)	43704-44730: c23 [9106-8099] (e-154) 43704-44730: c19 [23345-24352] (e-154) 44250-44729: c40 [18096-18575] (e-76)
mt188	44772-47992	44772-47992: c23 [7867-4653] (12800) 44772-47992: c19 [24584-27798] (12800) 46565-47992: c40 [24073-22663] (1798)	44790-47992: c23 [7851-4653] (≈0) 44790-47992: c19 [24600-27798] (≈0) 44792-46195, 46683-47992: c40 [18832-17371, 23958-22663] (e-193) also matches: c31 (e-15), c25 (e-8)
mt157	48098-48580	48098-48580: c23 [4494-4017] (1846) 48098-48580: c19 [27957-28494] (1846) 48098-48580: c40 [22504-22030] (949)	48105-48580: c23 [4488-4017] (e-94) 48105-48580: c19 [27963-28434] (e-94) 48180-48580: c40 [22055-22030] (e-54)

APPENDIX IVB: Results obtained from LOCALFASTA and TBLASTX searches performed on the contig sequences of *S. shibatae* against sequences from *S. solfataricus* P2.

Contig name	Position of contig on map	LOCAL FASTA searches against *S. solfataricus* P2 database	TBLASTX searches against *S. solfataricus* P2 database
shi113	1-1980		
shi126	2133-6750	2627-6750: c23 [46873-42747] (13741)	2647-6750: c23 [46852-42747] (≈0)
shi154	6857-7375	6857-7374: c23 [42471-41955] (1739)	6857-7375: c23 [42471-41954] (e-111)
shi2	7412-7915	7412-7915: c23 [41802-41300] (1669)	7412-7915: c23 [41802-41300] (e-99)
shi102	8000-13101	8013-11789: c23 [41207-37446] (12235) also matches : c20 (824) and c21 (628)	8061-13100: c23 [41164-35534] (≈0) also matches: c20 (e-24), c21 (e-23), c16 (e-12), c10 (e-12), c18 (e-10), c42 (e-10), c32 (e-9), c27 (e-9), c2 (e-8) ...
shi129	13195-14040	13196-14040: c23 [35304-34462] (3178)	13195-14040: c23 [35305-34462] (e-188)
shi200	14054-14888	14054-14888: c23 [34466-33651] (2896)	14136-14888: c23 [34142-33651] (e-161)
shi85	15326-15785	15326-15785: c23 [33083-32625] (1635)	15326-15785: c23 [33083-32626] (e-82)
shi116	16500-17244	16500-17244: c23 [31807-31065] (2846) 16500-17244: c19 [644-1385] (2846)	16500-17244: c23 [31807-31066] (e-158) 16500-17244: c19 [644-1385] (e-158)
shi14	17662-19148	17662-19148: c23 [30611-29128] (5551) 17662-19147: c19 [1841-3323] (5551)	17676-19148: c23 [30596-29127] (e-319) 17676-19148: c19 [1855-3324] (e-317) 17748-18509: c18 [12874-13641] (e-54)
shi162	19230-21398	19230-21398: c23 [29012-26832] (7891) 19230-21398: c19 [3439-5619] (7891)	19230-21398: c23 [29012-26832] (≈0) 19230-21398: c19 [3439-5619] (≈0)
shi177	21450-25903	21450-25233: c23 [26718-22941] (13444) 21450-25236: c19 [5733-9512] (13444)	21493-25903: c23 [26681-22295] (≈0) 21493-25903: c19 [5770-10156] (≈0) 24739-24963: c16 [16671-16895] (e-10)
shi61	26002-28394	26002-28130: c23 [22194-20080] (6459) 26002-28131: c19 [10257-12371] (6459) also matches: c22 (477), c52 (410) and c14 (410)	26012-28394: c23 [22184-19909] (≈0) 26012-28394: c19 [10267-12542] (≈0) also matches: c52 (e-13), c14 (e-13), c22 (e-12), c24 (e-11), c13 (e-10)
shi21	28500-29419	28500-29280: c23 [19697-18924] (2671) 28500-29281: c19 [12754-13527] (2671)	28585-29282: c23 [19620-18923] (e-146) 28585-29282: c19 [12831-13528] (e-146)
shi57	30010-30452	30010-30452: c23 [17959-17518] (1747) 30010-30452: c19 [14492-14933] (1747)	30010-30452: c23 [17959-17518] (e-104) 30010-30452: c19 [14492-14933] (e-104)
shi140	30500-30957	30500-30957: c23 [17470-17016] (1753) 30500-30957: c19 [14981-15435] (1753)	30500-30957: c23 [17426-17016] (e-101) 30500-30957: c19 [14981-15435] (e-101)
shi118	31813-32552	31813-32552: c23 [16657-15919] (2872) 31813-32552: c19 [15794-16532] (2872)	31813-32552: c23 [16657-15919] (e-169) 31813-32552: c19 [15794-16532] (e-169)
shi207	32636-34151	32636-34151: c23 [15563-14049] (5948) 32636-34151: c19 [16888-18402] (5948)	32636-34151: c23 [15563-14049] (≈0) 32636-34151: c19 [16888-18402] (≈0)
shi28	34383-35450	34383-34618: c23 [13773-13552] (665) 34383-34619: c19 [18678-18899] (665)	34463-34640: c23 [13707-13530] (e-7) 34463-34640: c19 [18744-18921] (e-7)
shi32	35855-36273		36129-36196: c23 [13254-13187] (e-4) 36129-36196: c19 [19197-19264] (e-4)
shi134	36645-37644	36645-37644: c23 [5061-4064] (3161) 36645-37644: c19 [27390-28387] (3161) 36645-37644: c40 [23071-22077] (1953)	36645-37644: c23 [5061-4064] (e-189) 36645-37644: c19 [27390-28387] (e-189) 36645-37644: c40 [23071-22077] (e-162)
shi69	37695-41110	37695-41110: c23 [3998-587] (12758)	37695-41110: c23 [3998-587] (≈0)

		37695-41110: c19 [28453-31864] (12758) also matches: c40 (1281), c42 (682)	37695-41110: c19 [28453-31864] (≈0) also matches: c40 (e-122). c25 (e-78), c42 (e-77), c1 (e-29), c16 (e-29), c29 (e-28), c10 (e-27), c18 (e-24) ...
shi186	41138-41822	41138-41692: c23 [550-1] (1854) 41138-41822: c19 [31901-32579] (2321)	41179-41692: c23 [515-2] (e-102) 41179-41822: c19 [31936-32579] (e-131)
shi204	41900-42713	41900-42713: c19 [32660-33477] (2918)	41920-42713: c19 [32680-33473] (e-180)
shi13	43016-44062	43294-44062: c19 [33816-34575] (2463)	43032-43106, 43318-44062: c19 [33750-33822, 33838-34575] (e-141)

APPENDIX IVC: Results obtained from LOCALFASTA and TBLASTX searches performed on the contig sequences of *S. acidocaldarius* against sequences from *S. solfataricus* P2.

Contig name	Position of contig on map	LOCAL FASTA searches against *S. solfataricus* P2 database	TBLASTX searches against *S. solfataricus* P2 database
a91	1-506		
a19	598-977		
a29	1610-3212		
a112	3290-3900		
a151	3960-4363		
a191	4438-4924		
a69	5155-6488		6312-6444: c23 [46853-46721] (e-11)
a110	6513-8793	7362-8793: c23 [45754-44338] (1231)	6602-8793: c23 [46567-44338] (e-246)
a111	8890-10208	8890-10186: c23 [44231-42925] (1674)	8890-10208: c23 [44232-42904] (e-140)
a88	10277-10770	10316-10765: c23 [42909-42466] (578)	10360-10766: c23 [42869-42465] (e-46)
a43	11077-12471	11077-12471: c23 [42213-40823] (1079)	11077-12471: c23 [42214-40938] (e-153)
a107	12617-14024	12746-12998: c23 [40635-40380] (471)	12620-12998, 13183-14023: c23 [40739-40380, 36508-35660] (e-91)
a18	14129-15461		14151-15456: c23 [35530-32337] (e-80) 15343-15451: c19 [5-118] (e-12)
a78	15643-17491	15643-16145: c23 [32203-31709] (836) 15643-16143: c19 [248-739] (836)	15643-16076, 16103-17001: c23 [32203-31772, 29677-28778] (e-174) 15643-16084, 16102-17001: c19 [248-687, 2773-3673] (e-175)
a146	17573-18467		17584-18007: c27 [20841-21270] (e-24) also matches: c32 (e-19), c5 (e-17), c25 (e-15), c1 (e-14), c23 (e-23), c30 (e-12), c29 (e-12), c57 (e-11), c19 (e-11)
a198	18655-19174		
a182	19195-20577	20065-20577: c23 [18571-18056] (1385) 20066-20577: c19 [13880-14394] (1385)	20137-20571: c23 [18501-18061] (e-66) 20137-20571: c19 [13950-14390] (e-66)
a108	22088-22695	22088-22695: c23 [16544-15937] (1772) 22088-22695: c19 [15907-16514] (1772)	22088-22695: c23 [16544-15937] (e-89) 22088-22695: c19 [15907-16514] (e-89)
a27	23556-24009	23556-24009: c23 [15089-14635] (1613) 23556-24009: c19 [17362-17816] (1613)	23556-24009: c23 [15089-14635] (e-89) 23556-24009: c19 [17362-17816] (e-89)
a147	24283-24702	24283-24702: c23 [14360-13942] (1268) 24283-24702: c19 [18091-18510] (1268)	24283-24697: c23 [14360-13947] (e-66) 24283-24693: c19 [18091-18500] (e-66)
a160	25283-28812		
a116	29008-31218		
a71	31390-37477		32455-33063: c57 [2535-3137] (e-22)

APPENDIX IVD: Results obtained from LOCALFASTA and TBLASTX searches performed on the contig sequences of *S. tokodaii* against sequences from *S. solfataricus* P2.

Contig name	Position of contig on map	LOCAL FASTA searches against *S. solfataricus* P2 database	TBLASTX searches against *S. solfataricus* P2 database
t13	1-225		
t2	830-1421		
t29	1831-4927		
t158	5012-5459	5060-5459: c22 [14052-14435] (908)	5088-5376: c22 [14078-14361] (e-45)
t55	5610-6983	5610-5791: c22 [14726-14905] (446)	5621-5749: c24 [14828-14700] (e-14) also matches: c22 (e-13)
t153	7062-7546	7073-7515: c39 [32382-31938] (441) also matches: c40 (441)	7154-7510: c39 [32308-31943] (e-36) also matches: c40 (e-36)
t176	7630-8779		7677-8057: c39 [31662-31291] (e-13) also matches: c40 (e-36)
t103	8809-10557		
t102	11480-12103		
t200	12153-12880		
t11	13307-13769	13536-13768: c24 [13069-13295] (504)	13550-13768: c24 [13080-13294] (e-29)
t219	13887-16027	15190-16027: c6 [26549-25715] (1538)	14109-14936, 15155-16026: c6 [29567-30525, 26585-25716] (e-153) 14109-14936: ISC1228orf [169-927] (e-25) also matches: c22 (e-14)
t73	16206-17056	16206-17056: c6 [25629-24774] (1442)	16207-17056: c6 [25628-24774] (e-139)
t39	17401-18858	17401-17930: c6 [24616-24090] (1062)	17436-17934: c6 [24584-24086] (e-93)
t109	18928-28195	19199-20658: c51 [22393-20935] (2958) also matches: c38 (2958) 24291-25465: c13 [2536-3721] (3404) also matches: c39 (3088), c40 (3088), c17 (2323), c18 (2210), c42 (2210), c25 (1983)...	19190-20546: c51 [22405-21047] (e-251) also matches: c38 (e-251) 24291-25385: c13 [2537-3645] (e-183) also matches: c17 (e-168), c39 (e-154), c40 (e-152), c18 (e-128), c42 (e-121), c25 (e-115)
t175	28315-31416	28951-29192: c13 [3719-3480] (607) also matches: c17 (532), c25 (572), c18 (575), c15 (547), c42 (566), c40(559)...	28927-29193: c25 [38662-38924] (e-15) also matches: c18 (e-23), c42 (e-20), c17 (e-18), c13 (e-17), c39 (e-15), c40 (e-15)
t58	31478-32530	31857-32530: c23 [18889-18211] (809) 32242-32530: c19 [13950-14240] (809)	32235-32479: c23 [18507-18257] (e-22) 32235-32479: c19 [13944-14194] (e-22)
t91	32642-33084	32642-33084: c23 [18114-17671] (1394) 32642-33084: c19 [14337-14780] (1394)	32642-33084: c23 [18114-17671] (e-69) 32642-33084: c19 [14337-14780] (e-69)
t76	34194-34633	34194-34633: c23 [16511-16073] (1305) 34194-34633: c19 [15940-16437] (1305)	34194-34633: c23 [16512-16074] (e-66) 34194-34633: c19 [15939-16377] (e-66)
t59	34952-38504	34952-37165: c23 [15795-13598] (6318) 34952-37216: c19 [16657-18904] (6318)	34997-37175: c23 [15758-13568] (≈0) 34996-37165: c19 [16692-18852] (≈0)
t119	38544-39631		
t214	39652-40379		
t9	40574-41474		
t46	41643-42474		41704-42444: c42 [16612-17427] (e-32) also matches: c31 (e-31), c57 (e-12), c25 (e-9)
t146	42580-43333		
t95	43485-43945		
t208	44314-45175		
t229	46230-46743		

APPENDIX IVE: Results obtained from LOCALFASTA and TBLASTX searches performed on the contig sequences of *M. sedula* against sequences from *S. solfataricus* P2.

Contig name	Position of contig on map	LOCAL FASTA searches against *S. solfataricus* P2 database	TBLASTX searches against *S. solfataricus* P2 database
m228	1-533		
m143	561-1952		
m105	2124-2551		
m294	2651-4126		
m206	4387-5128		
m320	5171-7006		
m47	10006-21699	17778-18708: c23 [45134-44199] (949)	16036-21678: c23 [46852-41211] (≈0)
m84	21717-22135		
m232	22145-23381	22642-22836: c23 [40478-40380] (641)	22369-22839, 22976-23304: c23 [40835-40378, 34484-34158] (e-72)
m192	23382-26182	25515-26182: c23 [31713-32379] (661) 25515-26182: c19 [738-72] (661)	24979-25545, 25578-26117: c23 [29112-29684, 31770-32309] (e-147) 24979-25545, 25578-26117: c19 [3339-2767, 681-142] (e-147)
m265	26299-38131	27531-29955: c31 [11579-13999] (4080) 30041-31192: c24 [7998-9148] (1453) 37502-38131: c23 [18723-18092] (1310) 37658-38131: c19 [13893-14359] (1310)	27532-33500: c31 [11581-15759] (≈0) 36428-37450: c15 [40000-41022] (e-59) 37708-38131: c23 [18501-18092] (e-61) 37708-38131: c19 [13950-14359] (e-61)
m21	41131-41822	41131-41822: c23 [15246-14559] (2241) 41131-41822: c19 [17205-17891] (2241)	41131-41807: c23 [15246-14573] (e-104) 41131-41807: c19 [17205-17878] (e-104)
m114	41988-46265	41988-42821: c23 [14391-13551] (1999) 41988-42822: c19 [18060-18900] (1999)	41989-42679: c23 [14391-13699] (e-86) 41989-42679: c19 [18061-18752] (e-85) 45537-45715: c2 [2423-2601] (e-14)
m255	46715-48297		
m70	48302-49134		
m69	50702-51914		51270-51632: c97 [34255-34290] (e-11)
m38	52369-54829		54496-54786: c57 [2535-2825] (e-12)
m268	54837-55398		54904-55308: c57 [2940-3356] (e-12)
m183	55459-59259		

APPENDIX IVF: Results obtained from LOCALFASTA and TBLASTX searches performed on the contig sequences of *S. solfataricus* P1 against sequences from *S. solfataricus* P2.

Contig name	Length of contig (bp)	LOCAL FASTA searches against *S. solfataricus* P2 database	TBLASTX searches against *S. solfataricus* P2 database
1	494	1-494: c23 [32623-32135] (1890) 174-494: c19 [1-316] (1198)	2-470: c23 [32622-32154] (e-100) 174-470: c19 [1-297] (e-59)
10	452	1-452: c23 [8635-9081] (1714) 1-452: c19 [23816-23370] (1714)	1-452: c23 [8635-9081] (e-73) 1-452: c19 [23816-23370] (e-73)
11	440	1-440: c23 [2767-3205] (1736) 1-440: c19 [29684-29246] (1736)	1-440: c23 [2767-3205] (e-90) 1-440: c19 [29684-29246] (e-90)
13	476	1-476: c23 [30290-30763] (1844) 1-476: c19 [2161-1688] (1844)	1-461: c23 [30290-30750] (e-98) 1-461: c19 [2161-1701] (e-98)
14	439	1-439: c23 [31788-31357] (1637) 1-439: c19 [663-1094] (1637)	1-439: c23 [31788-31357] (e-78) 1-439: c19 [663-1094] (e-78)
15	454	1-454: c19 [35697-35244] (1793)	1-454: c19 [35697-35244] (e-100)
17	981	1-981: c23 [4346-3369] (3862) 1-981: c19 [28105-29082] (3862)	1-981: c23 [4346-3369] (e-214) 1-981: c19 [28105-29082] (e-214)
19	427	1-427: c23 [7066-7490] (1676) 1-427: c19 [25385-24961] (1676)	1-427: c23 [7066-7490] (e-85) 1-427: c19 [25385-24961] (e-85)
20	406	1-406: c23 [8199-7798] (1513) 1-406: c19 [24252-24652] (1513)	1-394: c23 [8199-7810] (e-83) 1-394: c19 [24252-24641] (e-83)
21	541	1-541: c23 [17171-17713] (2121) 1-541: c19 [15280-14738] (2121)	1-541: c23 [17171-17713] (e-116) 1-541: c19 [15280-14738] (e-117)
22	398	1-398: c23 [14342-14735] (1455) 1-398: c19 [18109-17716] (1455)	1-374: c23 [14342-14714] (e-71) 1-374: c19 [18109-17737] (e-71)
23	437	1-437: c23 [16699-17132] (1640) 1-437: c19 [15752-15319] (1640)	1-437: c23 [16699-17132] (e-89) 1-437: c19 [15752-15319] (e-89)
25	1379	1-1379: c23 [25551-24173] (5488) 1-1379: c19 [6900-8278] (5488)	1-1379: c23 [25551-24173] (0) 1-1379: c19 [6900-8278] (0)
27	397	1-397: c23 [23476-23086] (1486) 1-397: c19 [8975-9365] (1486)	1-397: c23 [23476-23086] (e-77) 1-397: c19 [8975-9365] (e-77)
28	550	1-550: c23 [19457-18915] (2074) 1-550: c19 [12994-13536] (2074)	3-542: c23 [19455-18920] (e-104) 3-542: c19 [12996-13531] (e-101)
30	713	1-713: c19 [33006-33709] (2708)	1-652: c19 [33006-33656] (e-147)
31	916	1-916: c23 [5310-4401] (3568) 1-916: c19 [27141-28050] (3568)	1-911: c23 [5310-4405] (e-192) 1-911: c19 [27141-28046] (e-192)
34	422	1-422: c23 [1608-2026] (1619) 1-422: c19 [30843-30425] (1619)	1-401: c23 [1608-2007] (e-86) 1-401: c19 [30843-30444] (e-86)

35	491	1-491: c23 [29757-29269] (1964) 1-491: c19 [2694-3184] (1964)	1-491: c23 [29757-29269] (e-110) 1-491: c19 [2694-3184] (e-110)
38	1160	1-1160: c23 [26496-27639] (4358) 1-1160: c19 [5955-4812] (4358)	1-1160: c23 [26496-27639] (e-219) 1-1160: c19 [5955-4812] (e-219)
39	447	1-447: c23 [588-143] (1757) 1-447: c19 [31863-32308] (1757)	1-447: c23 [588-143] (e-98) 1-447: c19 [31863-32308] (e-98)
4	478	1-478: c23 [22211-22689] (1900) 1-478: c19 [10240-9762] (1900)	1-460: c23 [22211-22670] (e-109) 1-460: c19 [10240-9781] (e-109)
40	785	1-785: c19 [34694-33910] (3126)	1-785: c19 [34694-33910] (e-182)
43	436	1-436: c23 [9696-9267] (1673) 1-436: c19 [22753-23184] (1673)	2-436: c23 [9697-9267] (e-78) 2-436: c19 [22754-23184] (e-78)
44	410	1-410: c23 [5907-6310] (1545) 1-410: c19 [26544-26141] (1545)	1-337: c23 [5907-6243] (e-75) 1-337: c19 [26544-26208] (e-75)
45	402	1-402: c23 [21238-21636] (1560) 1-402: c19 [11213-10815] (1560)	1-395: c23 [21238-21632] (e-84) 1-395: c19 [11213-10819] (e-84)
6	455	1-455: c23 [20294-19840] (1792) 1-455: c19 [12157-12611] (1792)	1-455: c23 [20294-19840] (e-100) 1-455: c19 [12157-12611] (e-100)
8	785	1-785: c23 [28923-28142] (3085) 1-785: c19 [3528-4309] (3085)	1-766: c23 [28923-28159] (e-170) 1-766: c19 [3528-4291] (e-170)
9	430	1-430: c23 [10167-9742] (1640) 1-430: c19 [22284-22709] (1640)	1-415: c23 [10167-9755] (e-85) 1-415: c19 [22284-22695] (e-85)

APPENDIX VA: Results obtained from TBLASTX searches performed on the contig sequences of *S. solfataricus* MT4 against all contig sequences from *Sulfolobales*.

Contig name	Position of contig on map	TBLASTX searches against *S. shibatae* database	TBLASTX searches against *S. acidocaldarius* database	TBLASTX searches against *M. sedula* database	TBLASTX searches against *S. tokodaii* database
mt202	1-1334				10-182: t109 [27970-27801] (e-9)
mt165	1431-2782				
mt45	3137-3630	3137-3630: shi102 [12169-12658] (e-77)	3139-3624: a107 [13225-13707] (e-27)		
mt9	3875-4304	3875-4077: shi102 [12890-13091] (e-42)	3876-3981: a107 [13669-13773] (e-5)		
mt5	4334-4971	4334-4971: shi129 [13214-13850] (e-140)	4386-4949: a18 [14442-14693] (e-61)		
mt19	5157-5556	5251-5556: shi200 [14139-14443] (e-59)		5157-5481: m232 [22976-23304] (e-25)	
mt20	5606-6382	5616-6025: shi200 [14478-14888] (e-76)		5606-6362: m192 [23391-24115] (e-58)	
mt53	6389-7094	6528-6977: shi85 [15327-15783] (e-70)		6389-7043: m192 [24206-24843] (e-34)	
mt57	7139-8549	7799-8538: shi116 [16500-17244] (e-145)	7139-7269: a18 [15327-15456] (e-13) 7403-7834: a78 [15643-16076] (e-67)	7228-7836: m192 [26181-25578] (e-78)	
mt64	8571-10557	8960-10428: shi14 [17677-19148] (e-290)	9880-10509: a78 [16105-16740] (e-78)	9880-10488: m192 [25536-24940] (e-68)	
mt18	10721-14412	10721-11215: shi162 [19482-20094] (e-203) 13531-14412: shi162 [20252-21128] (e-203)			
mt91	14516-15557	14516-14625: shi162 [21265-21373] (e-9) 14809-15557: shi177 [21500-22247] (e-100)			
mt76	15626-15881	15628-15881: shi177 [22295-22553] (e-36)			
mt3	16883-17621	16883-17621: shi177 [23711-24440] (e-132)			
mt67	18635-19375	18635-19031: shi177 [25507-25903] (e-71)			
mt103	19716-20189				
mt17	20197-22184	20292-22184: shi61 [26034-27935] (≈0)			
mt87	22357-23596	23234-23396: shi61 [28232-28394] (e-10)			
mt101	23638-27007	23670-24366: shi21 [28585-29281] (e-142)			
mt182	27741-28053	27741-27800: shi57 [30392-30452] (e-7) 27848-28053: shi140 [30500-30706] (e-40)			
mt170	30621-31493	30622-31493: shi207 [33241-34110] (e-200)	30624-30947: a27 [23686-24009] (e-69) 31222-31492: a147 [24283-24552] (e-35)	30626-31009: m21 [41131-41807] (e-76) 31191-31493: m114 [41988-42289] (e-43)	30624-31493: t59 [35806-36674] (e-168)
mt195	31501-31928		31501-31637: a147 [24562-24697] (e-8)	31501-31885: m114 [42297-42679] (e-35)	31501-31877: t59 [36682-37059] (e-57)
mt179	32411-35630				35397-35580: t175 [30797-30983] (e-7)
mt98	35660-39485				
mt106	39488-39943				
mt183	40047-41616				
mt155	41987-43182				
mt185	43241-44731				
mt188	44772-47992	47583-47992: shi134 [36645-37054] (e-88)			
mt157	48098-48580	48144-48532: shi134 [37256-37644] (e-72)			

APPENDIX VB: Results obtained from TBLASTX searches performed on the contig sequences of *S. shibatae* against all contig sequences from *Sulfolobales*.

Contig name	Position of contig on map	TBLASTX searches against *S. solfataricus* MT4 database	TBLASTX searches against *S. acidocaldarius* database	TBLASTX searches against *M. sedula* database	TBLASTX searches against *S. tokodaii* database
shi113	1-1980		124-527: a151 [3967-4363] (e-42) 645-1028: a191 [4533-4915] (e-38) 1488-1920: a69 [5155-5625] (e-36)	194-1977: m47 [13621-15394] (e-194)	
shi126	2133-6750		2258-2763: a69 [5938-6429] (e-57) 2922-5155: a110 [6591-8793] (e-254) 5303-6592: a111 [8929-10208] (e-137) 6627-6740: a88 [10360-10473] (e-7)	2266-6750: m47 [15658-20128] (≈0)	
shi154	6857-7375		7114-7375: a43 [11077-11338] (e-77)	6859-7375: m47 [20423-20938] (e-77)	
shi2	7412-7915		7416-7911: a43 [11494-12016] (e-54)	7460-7914: m47 [21134-21588] (e-33)	
shi102	8000-13101	12169-12657: mt45 [3137-3629] (e-76) 12890-13092: mt9 [3875-4078] (e-41)	8078-8287: a43 [12170-12379] (e-22) 8486-8846: a107 [12620-12998] (e-96) 11164-11577: a146 [17573-17995] (e-6) 12132-12782: a107 [13186-13836] (e-96)	8147-8846: m232 [22145-22837] (e-48)	
shi129	13195-14040	13214-13850: mt5 [4334-4971] (e-139)	13265-13909: a18 [14442-15095] (e-62)		
shi200	14054-14888	14140-14443: mt19 [5252-5556] (e-59) 14478-14888: mt20 [5616-6025] (e-76)		14151-14370: m232 [23084-23304] (e-9) 14484-14788: m192 [23409-23705] (e-28)	
shi85	15326-15785	15326-15783: mt53 [6527-6977] (e-70)		15331-15709: m192 [24347-24685] (e-20)	
shi116	16500-17244	16500-17243: mt57 [7799-8537] (e-145)			
shi14	17662-19148	17677-19148: mt64 [8960-10428] (e-290)	18567-19145: a78 [16072-16656] (e-80)	18595-19145: m192 [25541-24997] (e-71)	
shi162	19230-21398	19409-21128: mt18 [11198-14412] (e-203) 21292-21373: mt91 [14544-14625] (e-8)			
shi177	21450-25903	21493-22247: mt91 [14802-15557] (e-99) 22295-22553: mt76 [15628-15881] (e-36) 23711-24440: mt3 [16883-17621] (e-131) 25507-25903: mt67 [18635-19031] (e-70)			
shi61	26002-28394	26034-27933: mt17 [20292-22182] (≈0) 28232-28394: mt87 [23234-23396] (e-10)			
shi21	28500-29419	28585-29281: mt101 [23670-24366] (e-145)			
shi57	30010-30452	30392-30452: mt182 [27741-27800] (e-6)			30010-30298: t91 [32801-33084] (e-58)
shi140	30500-30957	30500-30706: mt182 [27848-28053] (e-40)			
shi118	31813-32552		31927-32529: a108 [22088-22689] (e-98)		31959-32397: t76 [34194-34633] (e-71)
shi207	32636-34151	33241-34110: mt170 [30622-31493] (e-200)	33111-33563: a27 [23556-24007] (e-96) 33840-34151: a147 [24283-24595] (e-58)	32954-33627: m21 [41131-41807] (e-112) 33809-34151: m114 [41988-42330] (e-59)	32636-34151: t59 [35189-36715] (e-269)
shi28	34383-35450				
shi32	35855-36273				
shi134	36645-37644	36645-37054: mt188 [47583-47992] (e-89) 37256-37644: mt157 [48144-48532] (e-72)			
shi69	37695-41110	39943-40200: mt19 [5284-5541] (e-7)	39158-39703: a146 [18100-17573] (e-15)		
shi186	41138-41822				
shi204	41900-42713				
shi13	43016-44062				

APPENDIX VC: Results obtained from TBLASTX searches performed on the contig sequences of *S. acidocaldarius* against all contig sequences from *Sulfolobales*.

Contig number	Position of contig on map	TBLASTX searches against *S. solfataricus* MT4 database	TBLASTX searches against *S. shibatae* database	TBLASTX searches against *M. sedula* database	TBLASTX searches against *S. tokodaii* database
a91	1-506			95-470: m206 [5119-4747] (e-37)	
a19	598-977			603-762: m206 [4552-4393] (e-19) 822-976: m294 [4089-3936] (e-14)	
a29	1610-3212			1614-1949: m294 [3312-3013] (e-14) 2195-3212: m320 [6961-5968] (e-130)	
a112	3290-3900			3292-3400: m320 [5881-5773] (e-15) 3401-3821: m47 [12998-13421] (e-27)	
a151	3960-4363		3994-4363: shi113 [124-527] (e-43)	3994-4363: m47 [13587-13937] (e-31)	
a191	4438-4924		4533-4921: shi113 [645-1036] (e-39)	4443-4924: m47 [13950-14446] (e-45)	
a69	5155-6488		5155-5625: shi113 [1448-1920] (e-36) 5883-6444: shi126 [2200-2778] (e-58)	5253-6444: m47 [14949-16167] (e-146)	
a110	6513-8793		6532-8793: shi126 [2863-5155] (e-259)	6532-8793: m47 [16252-18572] (e-257)	
a111	8890-10208		8890-10208: shi126 [5261-6592] (e-141)	8891-10186: m47 [18662-19951] (e-136)	
a88	10277-10770		10360-10473: shi126 [6627-6740] (e-8)	10353-10758: m47 [19999-20419] (e-46)	
a43	11077-12471		11077-11338: shi154 [7114-7375] (e-37) 11494-12016: shi2 [7417-7911] (e-54) 12170-12379: shi102 [8078-8287] (e-22)	11077-11987: m47 [20678-21556] (e-102)	
a107	12617-14024	13225-13707: mt45 [3139-3624] (e-27) 13669-13773: mt9 [3876-3981] (e-5)	12620-13001: shi102 [8486-8849] (e-97) 13183-14023: shi102 [12129-12974] (e-97)	12620-13022: m232 [22465-22861] (e-25)	
a18	14129-15461	14442-15026: mt5 [4386-4961] (e-61) 15327-15456: mt57 [7139-7269] (e-13)	14442-15095: shi129 [13265-13909] (e-62)		
a78	15643-17491	15643-16078: mt57 [7403-7836] (e-67) 16087-16740: mt64 [9862-10509] (e-88)	16072-16656: shi14 [18567-19145] (e-79)	15643-16782: m192 [26011-24883] (e-154)	
a146	17573-18467		17573-18100: shi69 [39686-39158] (e-17) 17573-17876: shi102 [11164-11458] (e-8)		
a198	18655-19174				
a182	19195-20577			20109-20539: m265 [37680-38130] (e-65)	20075-20375: t58 [32177-32479] (e-27)
a108	22088-22695		22088-22689: shi118 [31927-32529] (e-98)		22120-22557: t76 [34194-34633] (e-70)
a27	23556-24009	23686-24009: mt170 [30624-30946] (e-69)	23556-24009: shi207 [33111-33565] (e-97)	23556-24008: m21 [41291-41744] (e-91)	23556-24009: t59 [35675-36129] (e-92)
a147	24283-24702	24283-24552: mt170 [31222-31492] (e-36) 24562-24697: mt195 [31501-31637] (e-8)	24283-24595: shi207 [33840-34151] (e-58)	24284-24700: m114 [42020-42435] (e-59)	24283-24693: t59 [36404-36813] (e-71)
a160	25283-28812			27339-28725: m255 [46717-48103] (e-247)	
a116	29008-31218			29012-29820: m70 [48331-49127] (e-94) 29930-30268: m84 [21718-22055] (e-27) 30362-30910: m192 [26090-25584] (e-11) 30605-30929: m69 [50702-51027] (e-34) 30973-31095: m38 [53022-53144] (e-5)	
a71	31390-37477			31395-32784: m38 [53433-54825] (e-196) 32804-33356: m268 [54837-55398] (e-79) 33417-37196: m183 [55459-59258] (≈0)	

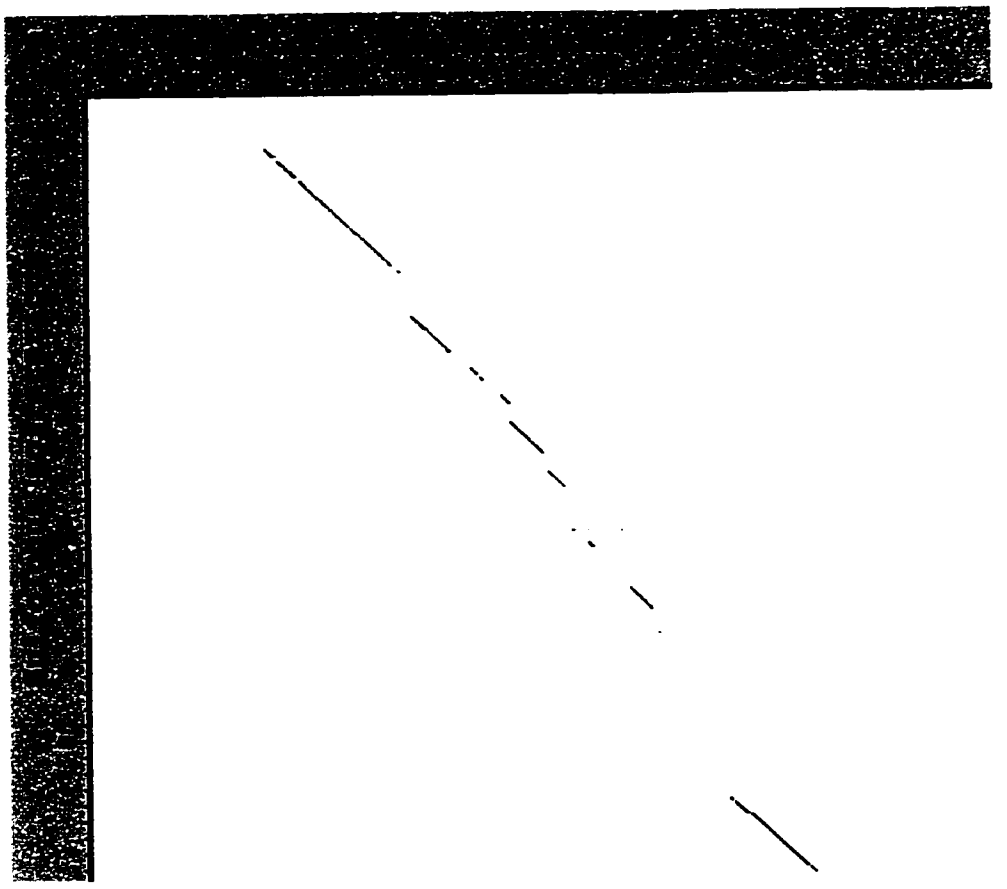
APPENDIX VD: Results obtained from TBLASTX searches performed on the contig sequences of *S. tokodaii* against all contig sequences from *Sulfolobales*.

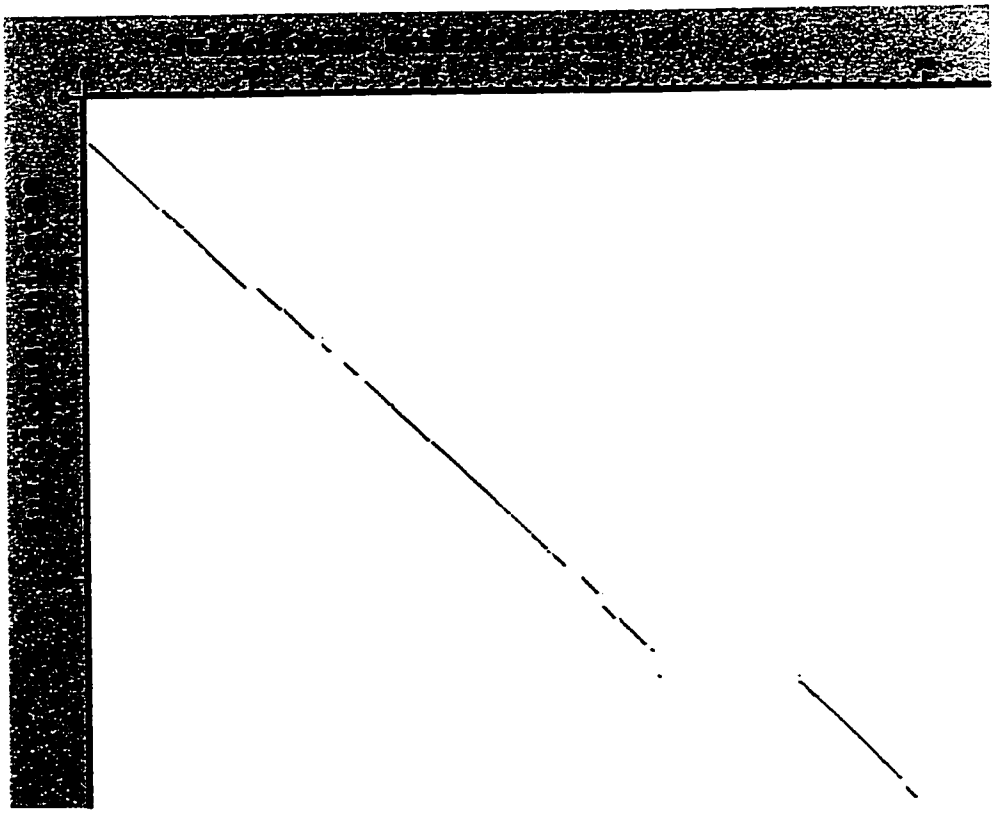
Contig number	Position of contig on map	TBLASTX searches against *S. solfataricus* MT4 database	TBLASTX searches against *S. shibatae* database	TBLASTX searches against *S. acidocaldarius* database	TBLASTX searches against *M. sedula* database
t13	1-225				
t2	830-1421				
t29	1831-4927				
t158	5012-5459				
t55	5610-6983				
t153	7062-7546				
t176	7630-8779				
t103	8809-10557				9980-10480: m47 [12911-12675] (e-30)
t102	11480-12103				
t200	12153-12880				
t11	13307-13769				
t219	13887-16027				
t73	16206-17056				
t39	17401-18858				
t109	18928-28195	27604-27968: mt202 [663-12] (e-8)			19821-20540: m265 [32980-33764] (e-19)
t175	28315-31416	30797-30983: mt179 [35395-35580] (e-7)			
t58	31478-32530			32182-32480: a182 [20080-20376] (e-27)	32241-32478: m265 [37708-37964] (e-24)
t91	32642-33084		32801-33084: shi57 [30010-30298] (e-58)		
t76	34194-34633		34194-34633: shi118 [31959-32397] (e-71)	34194-34633: a108 [22120-22557] (e-70)	
t59	34952-38504	35806-36674: mt170 [30624-31493] (e-168) 36682-37009: mt195 [31501-31829] (e-56)	35189-36715: shi207 [32636-34151] (e-269)	35675-36129: a27 [23556-24009] (e-90) 36404-36812: a147 [24283-24692] (e-73)	35519-36191: m21 [41131-41807] (e-111) 36373-37096: m114 [41988-42708] (e-85)
t119	38544-39631				
t214	39652-40379				
t9	40574-41474				
t46	41643-42474				
t146	42580-43333				
t95	43485-43945				
t208	44314-45175				
t229	46230-46743				

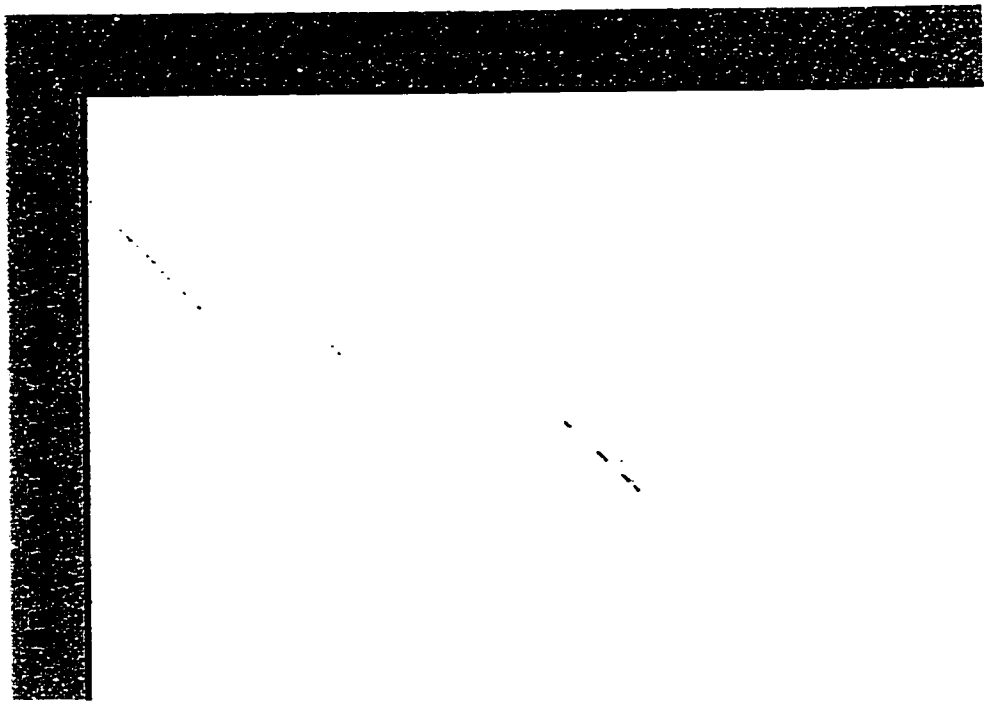
APPENDIX VE: Results obtained from TBLASTX searches performed on the contig sequences of *S. tokodaii* against all contig sequences from *Sulfolobales*.

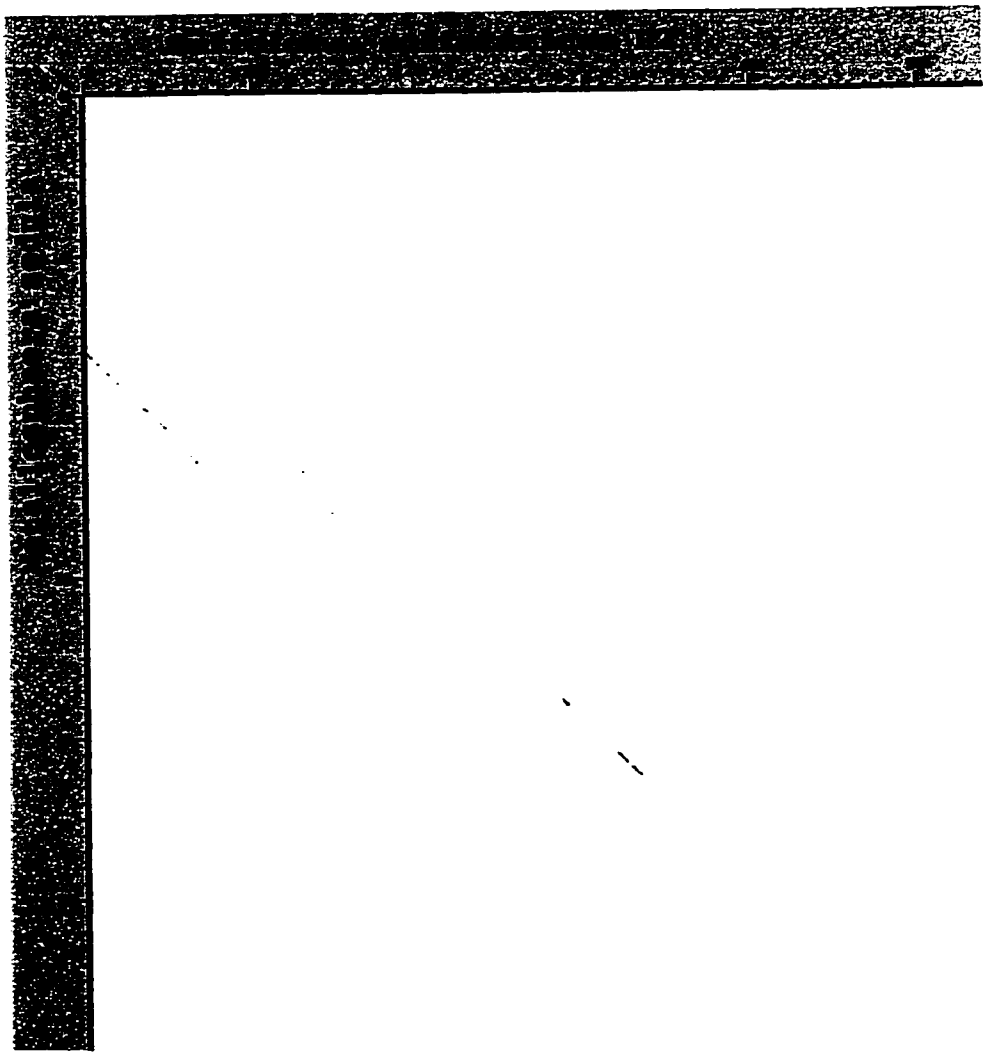
Contig number	Position of contig on map	TBLASTX searches against *S. solfataricus* MT4 database	TBLASTX searches against *S. shibatae* database	TBLASTX searches against *S. acidocaldarius* database	TBLASTX searches against *S. tokodaii* database
m228	1-533				
m143	561-1952				
m105	2124-2551				
m294	2651-4126			3013-3312: a29 [1949-1614] (e-14) 3936-4089: a19 [975-822] (e-13)	
m206	4387-5128			4393-4552: a19 [762-603] (e-19) 4846-5111: a91 [367-103] (e-37)	
m320	5171-7006			5968-7005: a29 [3212-2127] (e-130) 5773-5881: a112 [3400-3292] (e-15)	
m47	10006-21699		13452-15394: shi113 [16-1977] (e-193) 15658-19954: shi126 [2266-6574] (≈0) 20423-20938: shi154 [6859-7375] (e-76) 21134-21588: shi2 [7460-7914] (e-32)	12998-13421: a112 [3401-3821] (e-25) 13629-13937: a151 [4042-4363] (e-27) 14055-14445: a191 [4533-4923] (e-42) 14949-16167: a69 [5253-6444] (e-145) 16252-18572: a110 [6532-8793] (e-257) 18661-19970: a111 [8890-10205] (e-136) 19999-20419: a88 [10353-10758] (e-44) 20678-21696: a43 [11077-12127] (e-101)	12580-12911: t103 [10311-9980] (e-29)
m84	21717-22135			21718-22055: a116 [29930-30268] (e-28)	
m232	22145-23381	23036-23304: mt19 [5215-5481] (e-21)	22162-22837: shi102 [8164-8846] (e-49) 23084-23304: shi200 [14151-14370] (e-9)	22465-22836: a107 [12620-12997] (e-25)	
m192	23382-26182	23391-24115: mt20 [5606-6362] (e-58) 24206-24715: mt53 [6389-6890] (e-27) 24940-25536: mt64 [10488-9880] (e-68) 25578-26181: mt57 [7836-7228] (e-77)	23409-23705: shi200 [14486-14788] (e-27) 24347-24685: shi85 [15331-15709] (e-20) 24997-25539: shi14 [19145-18597] (e-71)	24883-26011: a78 [16782-15643] (e-154) 25584-26090: a116 [30910-30362] (e-11)	
m265	26299-38131	31642-32238: mt188 [45602-46183] (e-12)		37719-38130: a182 [20148-20539] (e-64)	32980-33764: t109 [19821-20540] (e-18) 37708-37964: t58 [32241-32478] (e-23)
m21	41131-41822	41423-41807: mt170 [30626-31009] (e-76)	41131-41807: shi207 [32954-33627] (e-113)	41291-41744: a27 [23556-24008] (e-91)	41131-41807: t59 [35519-36191] (e-112)
m114	41988-46265	41988-42289: mt170 [31191-31493] (e-43) 42296-42678: mt195 [31501-31885] (e-31)	41988-42330: shi207 [33809-34151] (e-59)	42020-42435: a147 [24284-24700] (e-58)	41988-42708: t59 [36373-37096] (e-85)
m255	46715-48297			46717-48105: a160 [27339-28727] (e-248)	
m70	48302-49134			48331-49127: a116 [29014-29820] (e-94)	
m69	50702-51914			50702-51027: a116 [30605-30929] (e-35)	
m38	52369-54829			53022-53144: a116 [30973-31095] (e-5) 53433-54823: a71 [31395-32782] (e-197)	
m268	54837-55398			54842-55398: a71 [32804-33356] (e-80)	
m183	55459-59259			55459-59258: a71 [33417-37197] (≈0)	

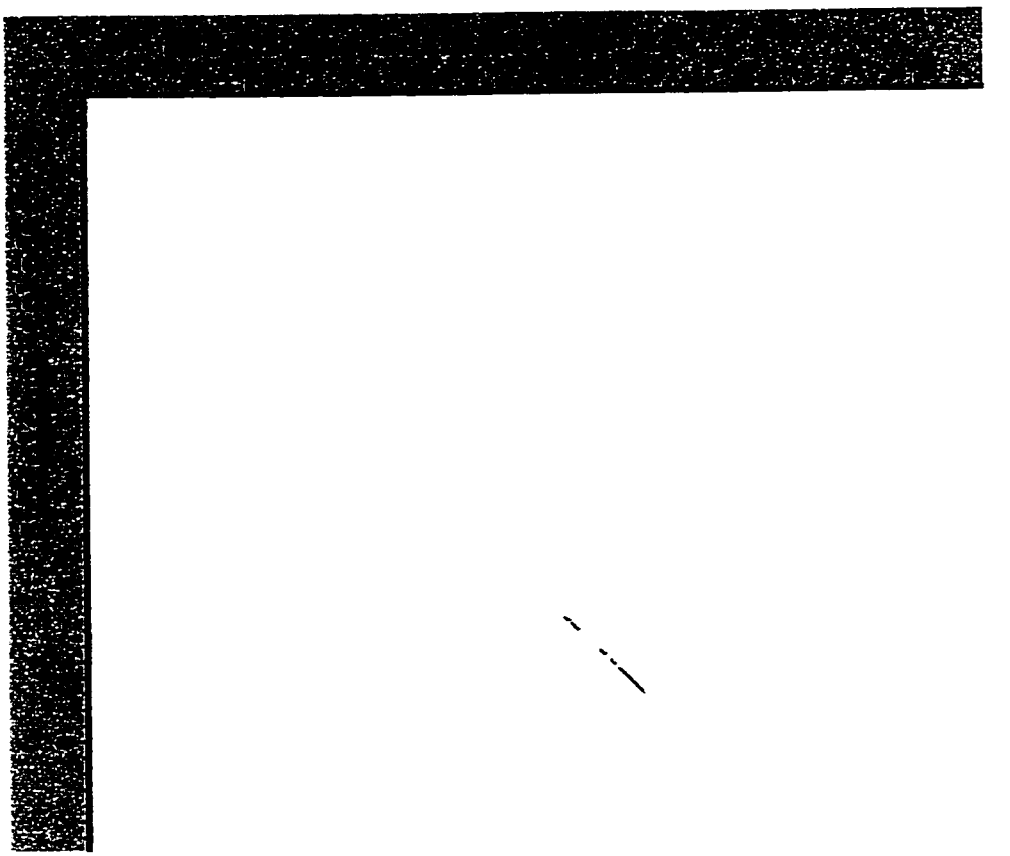
APPENDIX VI: Dot plots showing the similarities in gene organization between the different *Sulfolobales* at the amino acid level. Each dot plot has a 33 amino acids resolution and inserts a dot every time 5 amino acids out of 8 are matched. The intensity of the dot is indicative of the number of matches that are found within the 33 amino acid window.

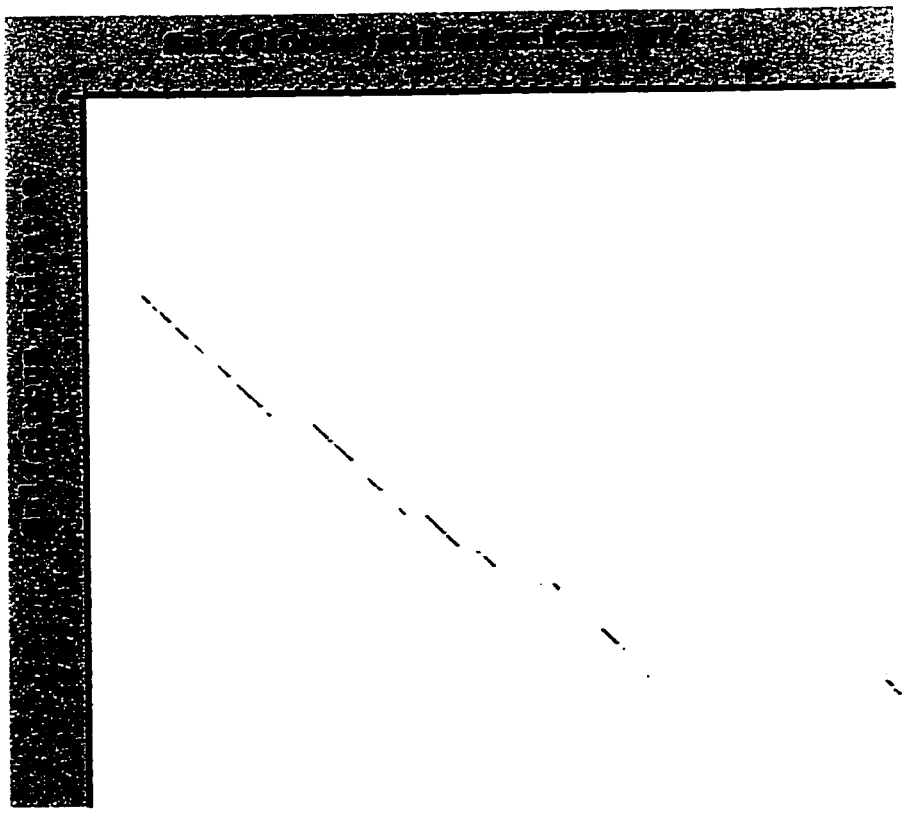


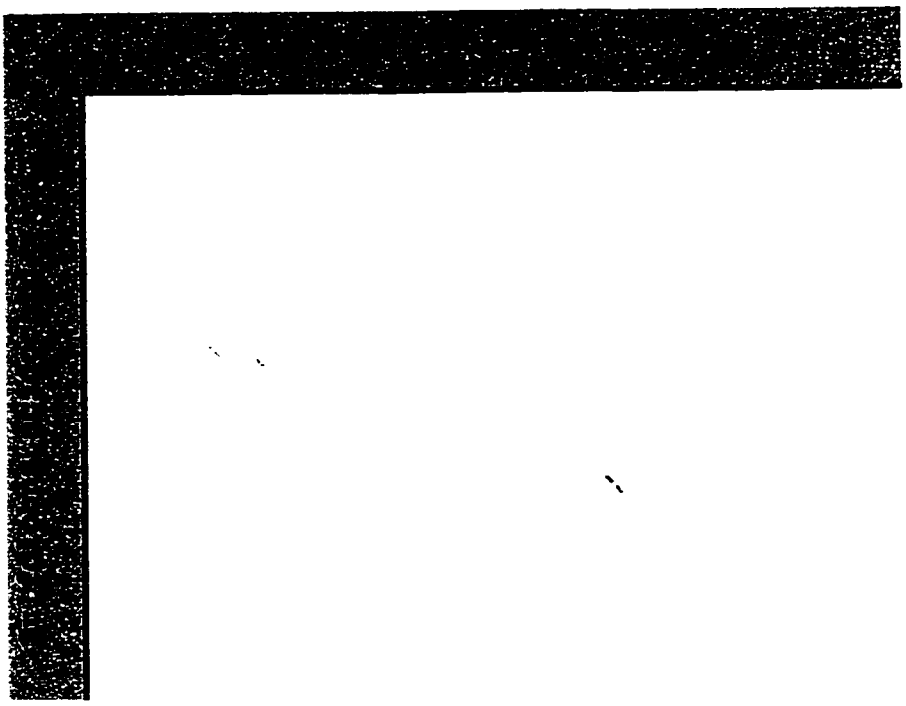


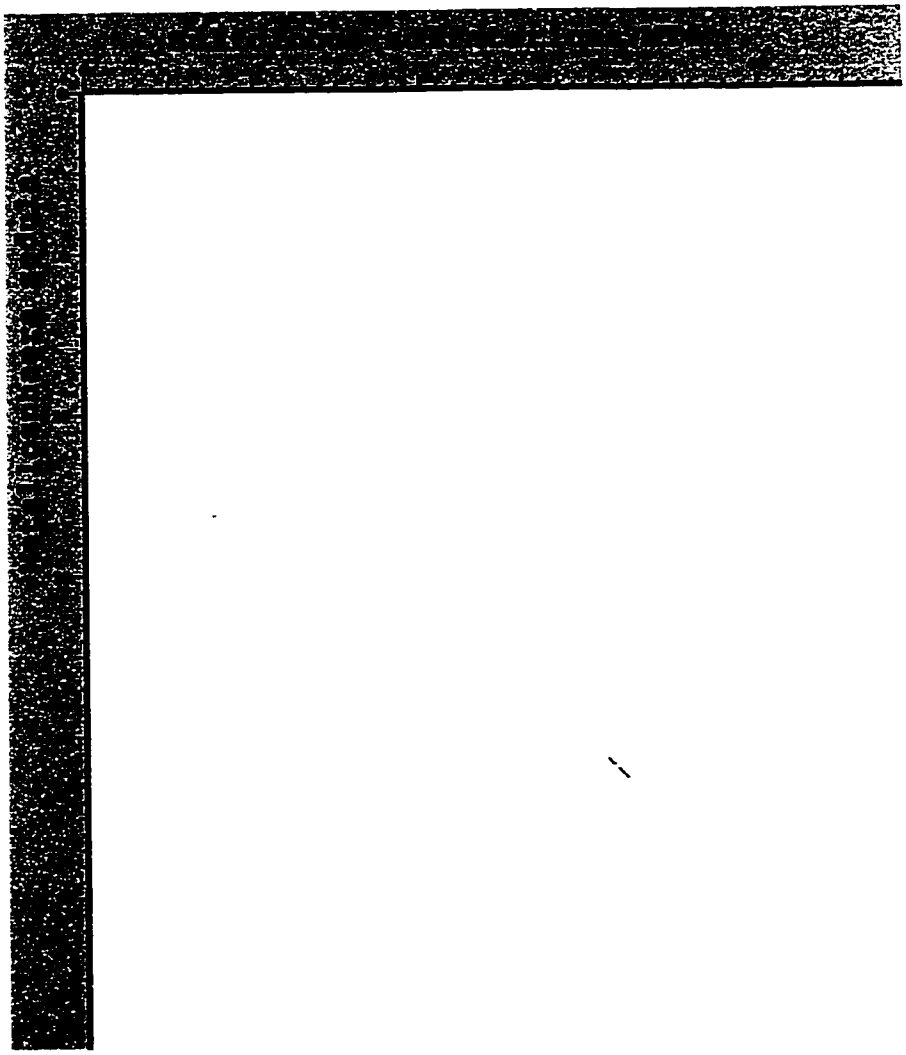


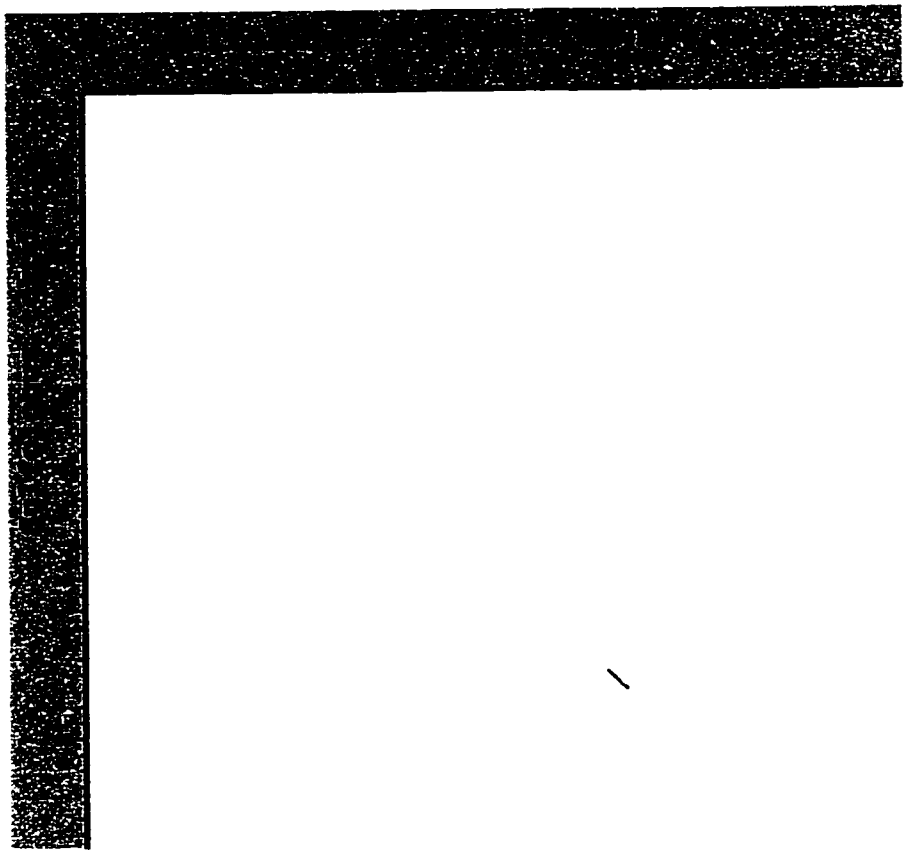


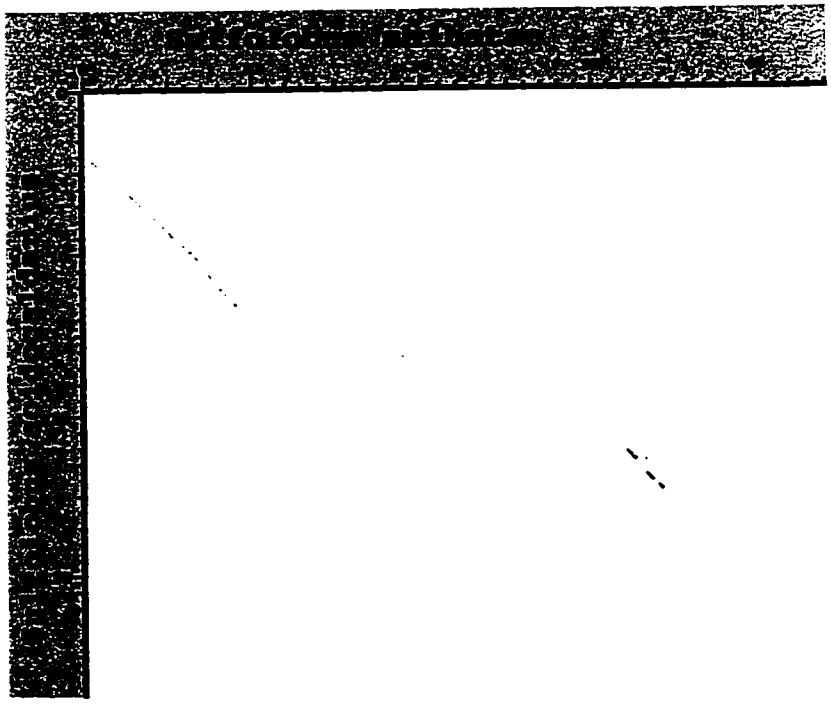


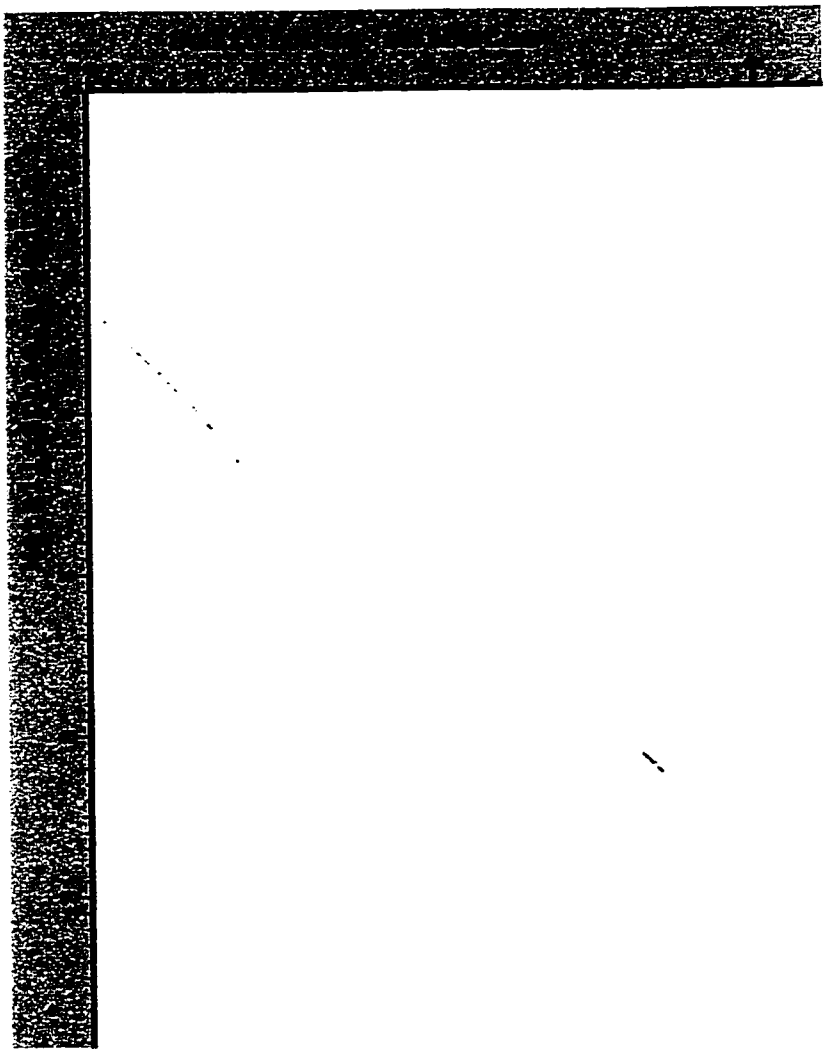


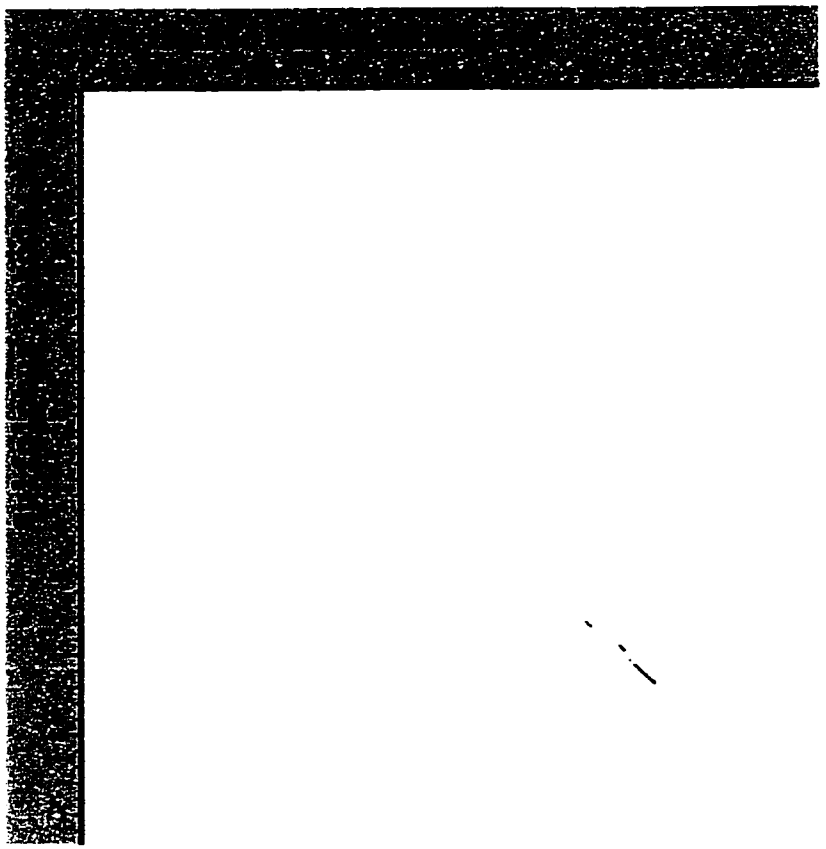


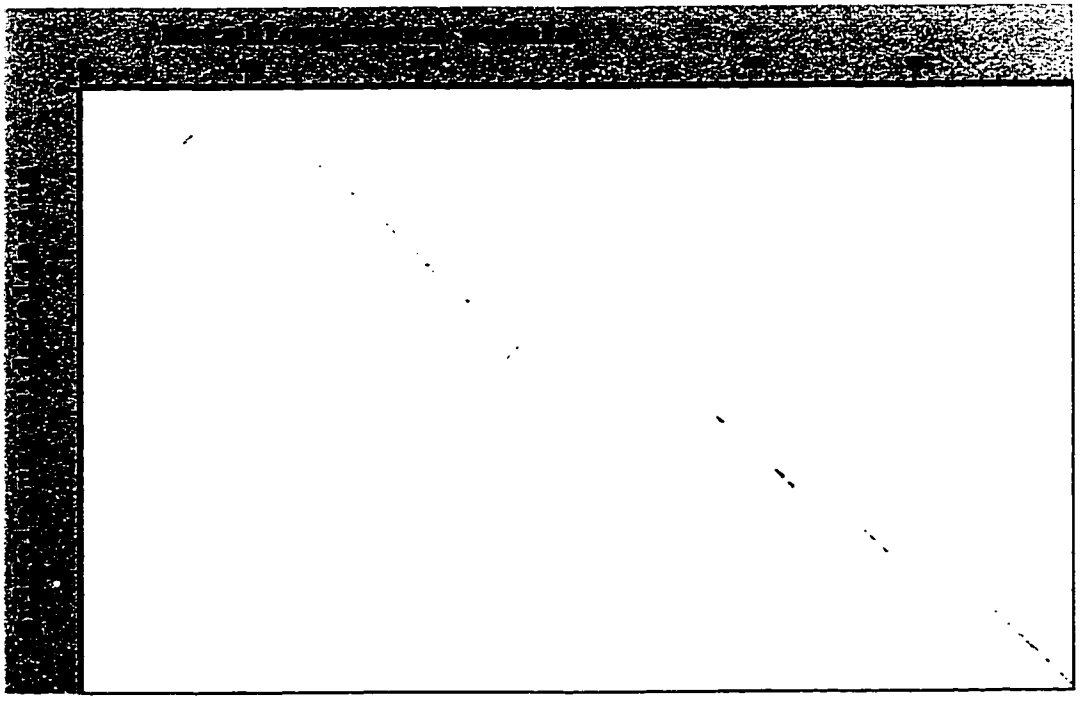


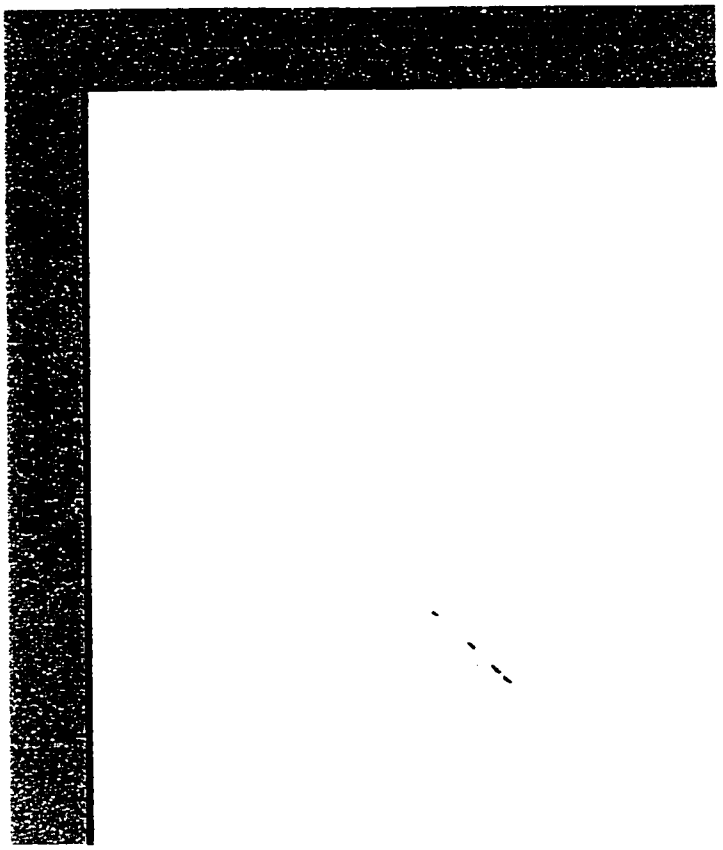


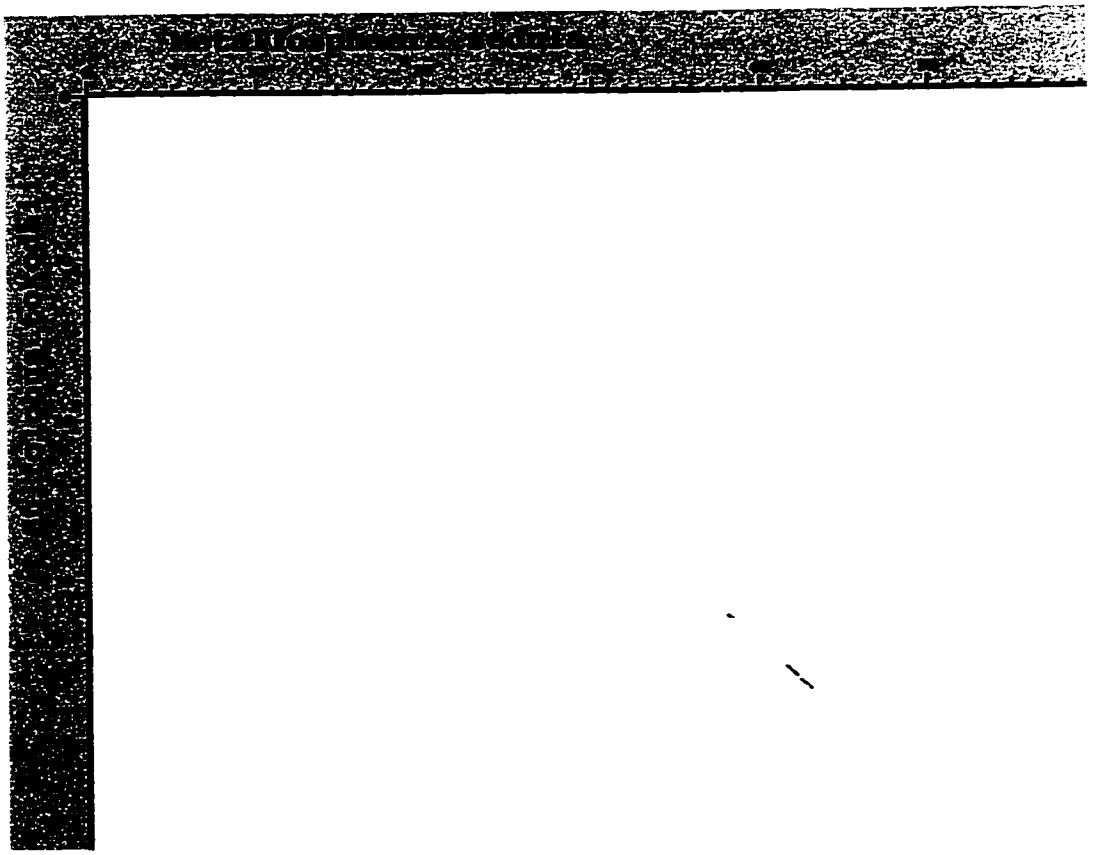












REFERENCES

Allen, M.B. 1959. Studies with *Cyanidium caldarium*, an anomalously pigmented chlorophyte. *Arch. Mikrobiol.* **32**:270-277.

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

Bachmann, B. 1983. Linkage map of *Escherichia coli* K-12, edition 7. *Microbiol. Res* **47**:180-230.

Balke, V.L., and J.D. Gralla. 1987. Changes in linking number of supercoiled DNA accompany growth transitions in *Escherichia coli*. *J. Bacteriology* **169**:4499-4506.

Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. and Y. Shao. 1997. The complete genome sequence of *Escherichia coli* K-12. *Science* **277**:1453-1474.

Brewer. B.J. 1990. Replication and the transcriptional organization of the *Escherichia coli* chromosome, p. 61-83. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

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

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.M., Weidman, J.F., Fuhrmann, J.L., Nguyen, D., Utterback, T.R., Kelley, J.M., Peterson, J.D., Sadow, P.W., Hanna, M.C., Cotton, M.D., Roberts, K.M., Hurst, M.A., Kaine, B.P., Borodovsky, M., Klenk, H.-P., Fraser, C.M., Smith, H.O., Woese, C.R., and J.C. Venter. 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*. **273**:1058-1073.

Canard, B., Saint-Joanis, B. and, S.T. Cole. 1992. Genomic diversity and organization of virulence genes in the pathogenic anaerobe *Clostridium perfringens*. *Mol. Microbiology*. **6**:1421-1429.

Carlson, C.R., and A.-B. Kolsto. 1993. A complete physical map of a *Bacillus thuringiensis* chromosome. *J. of Bacteriology*. **175**:1053-1060.

Casjens, S., Delange, M., Ley III, H.L., Rosa, P., and W.M. Huang. 1995. Linear chromosomes of Lyme disease agent spirochetes: Genetic diversity and conservation of gene order. *J. Bacteriology*. **177**:2769-2780.

Charlebois, R.L., and A. St. Jean. 1995. Supercoiling and map stability in the bacterial chromosome. *J. Mol. Evol.* **41**:15-23.

Chandler, M.S. 1997. Insertion sequences and transpositions, p.30-37. In Bruijn, F. J., Lupski, J.R. and G.M. Weinstock (ed.), Bacterial genomes Physical structure and analysis. Chapman & Hall, New York, NY.

Cole, S.T., and I. Saint Girons. 1994. Bacterial genomics. *FEMS Microbiology Reviews* **14**:139-160.

Dear, S., and R. Staden. 1991. A sequence assembly and editing program for efficient management of large sequencing projects. *Nucleic Acids Res.* **21**:2942.

Deckert, G., Warren, P.V., Gaasterland, T., Young, W.G., Lenox, A.L., Graham, D.E., Overbeek, R., Snead, M.A., Keller, M., Aujay, M., Huber, R., Feldman, R.A., Short, J.M., Olsen, G.J., and R. V. Swanson. 1998. The complete genome of the hyperthermophilic bacterium *Aquifex aeolicus*. *Nature* **392**: 353-358.

DeRosa, M., Gambacorta, A., and J.D. Bu'Lock. 1975. Extremely thermophilic acidophilic bacteria convergent with *Sulfolobus acidocaldarius*. *J. Gen. Microbiol.* **86**:156-164.

De Smet, K.A.L., Jamil. S., and N.G. Stoker. 1993. Tropist3: a cosmid vector for simplified mapping of both G+C-rich and A+T-rich genomic DNA. *Gene* **136**:215-219.

Dorman, C.J., Barr, G.C., Ni Bhriain, N. and C.F. Higgins. 1988. DNA supercoiling and the anaerobic and growth phase regulation of *tonB* gene expression. *J. Bacteriology* **170**:2816-2826.

Dorman, C.J. 1996. Flexible response: DNA supercoiling, transcription and bacterial adaptation to environmental stress. *Trends in Microbiology* **4**:214-216.

Drlica, K., Pruss, G.J., Burger, R.M., Franco, R.J., Hsieh, L.-S. and B.A. Berger. 1990. Roles of DNA topoisomerases in bacterial chromosome structure and function. p.195-204. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Dybvig, K. 1993. DNA rearrangements and phenotypic switching in prokaryotes. *Molecular Microbiology* **10**:465-471.

Ehrlich, S.D., Bierne, H., d'Alençon, E., Vilette, D., Petranovic, M., Noirot, P. and B. Michel. 1993. Mechanisms of illegitimate recombination. *Gene* **135**:161-166.

Felsenstein, J. 1993. PHYLIP (Phylogeny Inference Package) version 3.5c. Department of Genetics, University of Washington, Seattle.

Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.-F., Dougherty, B.A., Merrick, J.M., McKenney, K., Sutton, G., FitzHugh, W., Fields, C., Gocayne, J.D., Scott, J., Shirley, R., Liu, L.-I., Glodek, A., Kelley, J.M., Weidman, J.F., Phillips, C.A., Spriggs, T., Hedblom, E., Cotton, M.D., Utterback, T.R., Hanna, M.C., Nguyen, D.T., Saudek, D.M., Brandon, R.C., Fine, L.D., Fritchmann, J.L., Fuhrmann, J.L., Geoghagen, N.S.M., Gnehm, C.L., McDonald, L.A., Small, K.V., Fraser, C.M., Smith, H.O., and J.C. Venter. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science.* **269**:496-512.

Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.-F., Dougherty, B.A., Bott, K.F., Hu, P.-C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison III, C.A., and J.C. Venter. 1995. The minimal gene complement of *Mycoplasma genitalium. Science*, **270**:397-403.

Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R.A., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., Gwinn, M., Dougherty, B., Tomb, J.-F., Fleischmann, R.D., Richardson, D., Peterson, J., Kerlavage, A.R., Quackenbush, J., Salzberg, S., Hanson, M., van-Vugt, R., Palmer, N., Adams, M.D., Gocayne, J.D., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fujii, C., Cotton, M.D., Horst, K., Roberts, K., Hatch, B., Smith, H.O. and J.C. Venter. 1997. Genomic sequence of a Lyme disease spirochete. *Borrelia burgdorferi. Nature* **390**:580-586.

Fonstein, M. and R. Haselkorn. 1995. Physical mapping of bacterial genomes. *J. Bacteriol.* **177**:3361-3369.

Fonstein, M. and R. Haselkorn. 1997. Encyclopedias of bacterial genomes. p. 348-361. In Bruijn, F. J., Lupski, J.R. and G.M. Weinstock (ed.), Bacterial genomes Physical structure and analysis. Chapman & Hall, New York, NY.

François, V., Louarn, J., Rebello, J.-E. and J.-M. Louarn. 1990. Replication termination, nondivisible zones, and structure of the *Escherichia coli* chromosome, p.351-359. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Fuchs, T., Huber, H., Burggraf, S., and K.O. Stetter. 1996. 16S rDNA-based phylogeny of the archaeal order *Sulfolobales* and reclassification of *Desulfurolobus ambivalens* as *Acidianus ambivalens* comb. nov. *System. Appl. Microbiol.* **19**:56-60.

Ginard, M., Lalucat, J., Tümmler, B., and U. Rõmling. 1997. Genome organization of *Pseudomonas stutzeri* and resulting taxonomic and evolutionary considerations. *Int. J. Syst. Bacteriology.* **47**:132-143.

Grogan, D.W. 1989. Phenotypic characterization of the archaebacterial genus *Sulfolobus*: Comparison of five wild-type strains. *J. of Bacteriology.* **171**:6710-7719.

Grogan D., Palm, P., and W. Zillig. 1990. Isolate B12, which harbors a virus-like element represents a new species of the archaebacterial genus *Sulfolobus*, *Sulfolobus shibatae*, sp. nov.. *Arch. Microbiology.* **154**:594-599.

Hallet, B., and D.J. Sherratt. 1997. Transposition and site-specific recombination: adapting DNA cut-and-paste mechanisms to a variety of genetic rearrangements. *FEMS Microbiology Reviews* **21**:157-178.

Higgins, C.F., Dorman, C.J., Stirling, D.A., Waddell, L., Booth, I.R., May, G., and E. Bremer. 1988. A physiological role for DNA supercoiling in the osmotic control of gene expression in *S. typhimurium* and *E. coli*. *Cell* **52**:569-584.

Higgins, C.F., Dorman, C.J., and N. Ni Bhriain. 1990. Environmental influences on DNA supercoiling: a novel mechanism for the regulation of gene expression, p. 421-432. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Hill, C.W., Harvey, S., and J.A. Gray. 1990. Recombination between rRNA genes in *Escherichia coli* and *Salmonella typhimurium*, p.335-340. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.-C., and R. Herrmann. 1996. Complete sequence analysis of the genome of the bacterium *Mycoplasma pneumoniae*. *Nucleic Acids Res.* **24**:4420-4449.

Himmelreich, R., Plagens, H., Hilbert, H., Reiner, B., and R. Herrmann. 1997. Comparative analysis of the genomes of the bacteria *Mycoplasma pneumoniae* and *Mycoplasma genitalium*. *Nucleic Acids Res.* **25**:701-712.

Holloway, B.W., Dharmsthiti, S., Krishnapillai, V., Morgan, A., Obeyesekere, V., Ratnaningsih, E., Sinclair, M., Strom, D. and C. Zhang. 1990. Patterns of gene linkages in Pseudomonas species, p. 97-106. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Hopwood, D.A., and T. Kieser. 1990. The *Streptomyces* genome, p. 147-162. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Huber, G., Spinnler, C., Gambacorta, A., and K.O. Stetter. 1989. *Metallosphaera sedula* gen. and sp. nov. represents a new genus of aerobic, metal-mobilizing, thermoacidophilic archaebacteria. *System. Appl. Microbiology.* **12**:38-47.

Jiang, Q., Hiratsuka, K., and D.E. Taylor. 1996. Variability of gene order in different *Helicobacter pylori* strains contributes to genome diversity. *Mol. Microbiology.* **20**:833-842.

Jordi B.J., Owen-Hughes, T.A., Hulton, C.H., and C.F. Higgins. 1995. DNA twist, flexibility and transcription of the osmoregulated proU promoter of *Salmonella typhimurium*. *EMBO J.* **14**:5690-5700.

Jyothirmai, G. and R.K. Mishra. 1994. Differential influence of DNA supercoiling on in vivo strength of promoters varying in structure and organization in *E. coli*. *FEBS Letters.* **340**:189-192.

Kaneko, T., Sato., S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda., M., and S. Tabata. 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. *DNA Res.* **3**:109-136.

Keeling, P.J., Charlebois, R.L., and W.D. Doolittle. 1994. Archaebacterial genomes: eubacterial form and eukaryotic content. *Current Opinion in Gen. and Dev.* **4**:816-822.

Keeling, P.J. and F. Doolittle. 1995. Archaea: Narrowing the gap between prokaryotes and eukaryotes. *Proc. Natl. Acad. Sci. USA.* **92**:5761-5764.

Kellenberger, E. 1990. Intracellular organization of the bacterial genome. p. 173-186. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Klenk, H.P., Clayton, R.A., Tomb, J., 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., Peterson, S., Reich, C.I., McNeil, L.K., Badger, J.H., Glodek, A., Zhou, L., Overbeek, R., Gocayne, J.D., Weidman, J.F., McDonald, L., Utterback, T., Cotton, M.D., Spriggs, T., Artiach, P., Kaine, B.P., Sykes, S.M., Sadow, P.W., D'Andrea, K.P., Bowman, C., Fujii, C., Garland, S.A., Mason, T.M., Olsen, G.J., Fraser, C.M., Smith, H.O., Woese, C.R. and J.C. Venter. 1997. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*. **390**:364-370.

Kolsto, A.-B. 1997. Dynamic bacterial genome organization. *Mol. Microbiology* **24**:241-248.

Kondo, S., Yamagishi, A., and T. Oshima. 1993. A physical map of the Sulfur-dependent archaebacterium *Sulfolobus acidocaldarius* 7 chromosome. *J. of Bacteriology*. **175**:1532-1536.

Krawiec, S., and M. Riley. 1990. Organization of the bacterial chromosome. *Microbiological Reviews* **54**:502-539.

Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.K., Codani, J.J., Connerton, I.F., Danchin, A., *et al.* 1997. The complete genome sequence of the Gram-positive bacterium *Bacillus subtilis*. *Nature*. **390**:249-256.

Ladefoged, S.A., and G. Christiansen. 1992. Physical and genetic mapping of the genomes of five Mycoplasma hominis strains by pulsed-field gel electrophoresis. *J. Bacteriology*. **174**:2199-2207.

Lawrence J.G., and H. Ochman. 1997. Amelioration of bacterial genomes: Rates of change and exchange. *J. Mol. Evol.* **44**:383-397.

Leblond, P., Redenbach, M., and J. Cullum. 1993. Physical map of the *Streptomyces lividans* 66 genome and comparison with that of the related strain *Streptomyces coelicolor* A3(2). *J. Bacteriology*. **175**:3422-3429.

Le Bourgois, P., Lautier, M., Berghe, L.V.D., Gasson, M.J., and P. Ritzenthaler. 1995. Physical and genetic map of the *Lactococcus lactis* subsp. *cremoris* MG1363 chromosome: Comparison with that of *Lactococcus lactis* subsp. *lactis* IL 1403 reveals a large genome inversion. *J. Bacteriology*. **177**:2840-2850.

Liu, S.-L., Hessel, A. and K.E. Sanderson. 1993. The *Xba*I-*Bln*I-*Ceu*I genomic cleavage map of *Salmonella enteritidis* shows an inversion relative to *Salmonella typhimurium* LT2. *Mol. Microbiology.* **10**:655-664.

Liu, S.-L., and K.E. Sanderson. 1995a. I-CeuI reveals conservation of the genome of independent strains of *Salmonella typhimurium. J. Bacteriology.* **177**:3355-3357.

Liu, S.-L., and K.E. Sanderson. 1995b. Rearrangements in the genome of the bacterium *Salmonella typhi. Proc. Natl. Acad. Sci. USA.* **92**:1018-1022.

Liu, S.-L., and K.E. Sanderson. 1996. Highly plastic chromosomal organization in *Salmonella typhi. Proc. Natl. Acad. Sci. USA.* **93**:10303-10308.

López-García, P., St. Jean, A., Amils, R., and R.L. Charlebois. 1995. Genomic stability in the archaea *Haloferax volcanii* and *Haloferax mediterranei. J. Bacteriology.* **177**:1405-1408.

Lück, P.C., Helbig, J.H., Drasar, V., Bornstein, N., Fallon, R.J., and M. Castellani-Pastoris. 1995. Genomic heterogenicity amongst phenotypically similar *Legionella micdadei* strains. *FEMS Microbiology letters.* **126**:49-54.

Mahan, M.J., Segall, A.M., and J.R. Roth. 1990. Recombination events that rearrange the chromosome: Barriers to inversion, p. 341-349. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Matic, I., Rayssiguier, C., and M. Radman. 1995. Interspecies gene exchange in bacteria: The role of SOS and mismatch repair systems in evolution of species. *Cell* **80**:507-515.

McClelland, M., Wong, K.K., and K. Sanderson.1997. Physical mapping and fingerprinting of bacterial genomes using rare cutting restriction enzymes, p.253-311. In Bruijn, F. J., Lupski, J.R. and G.M. Weinstock (ed.), Bacterial genomes Physical structure and analysis. Chapman & Hall, New York, NY.

Mushegian, A.R., and E.V. Koonin. 1996. Gene order is not conserved in bacterial evolution. *Trends in genetics.* **12**:289-290.

Nickerson, C.A., and E.C., Achberger. 1995. Role of curved DNA in binding of *Escherichia coli* RNA polymerase to promoters. *J. of Bacteriology* **177**:5756-5761.

Nikolskaya, T., Fonstein, M., and R. Haselkorn. 1995. Alignment of a 1.2-Mb chromosomal region from three strains of *Rhodobacter capsulatus* reveals a significantly mosaic structure. *Proc. Natl. Sci. USA.* **92**:10609-10613.

Ochman, H., and A.C. Wilson. 1987. Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26**:74-86.

Ojaimai, C., Davidson, B.E., Saint Girons, I. and I.G. Old. 1994. Conservation of gene arrangement and an unusual organization of rRNA genes in the linear chromosomes of the Lyme disease spirochaetes *Borrelia burgdorferi, B. garinii* and *B. afzelii. Microbiology.* **140**:2931-2940.

Pattee, P.A. 1990. Genetic and physical mapping of the chromosome of *Staphylococcus aureus* NCTC 8325, p. 163-169. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Pace, N.R. 1996. New perspective on the natural microbial world: molecular microbial ecology. *ASM News.* **62**:463-470.

Pearson, W.R., and D. Lipman. 1988. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci.* USA. **85**:2444-2448.

Perkins, J.D., Heath, J.D., Sharma., B.R., and G.M. Weinstock. 1993. *Xba*I and *Bln*I genomic cleavage maps of Escherichia coli K-12 strain MG1655 and comparative analysis of other strains. *J. Mol. Biol.* **232**:419-445.

Piggot, P.J. 1990. Genetic map of *Bacillus subtilis* 168, p. 107-146. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Riley, M., and K.E. Sanderson. 1990. Comparative genetics of *Escherichia coli* and *Salmonella typhimurium*, p.85-96. In K. Drlica and M. Riley (ed.), The bacterial chromosome. American Society for Microbiology, Washington, D.C.

Sanderson, K., and J. Roth. 1988. Linkage map of *Salmonella typhimurium*, edition VII. *Microbiol. Rev.* **52**:485-532.

Sankoff, D., and M. Goldstein. 1989. Probabilistic models of genome shuffling. *Bull. Math. Biol.* **51**:117-124.

Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., and R. Cedergren. 1992. Gene order comparisons for phylogenetic inference: Evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. USA.* **89**:6575-6579.

Schmid, M.B., and J.R. Roth. 1987. Gene location affects expression level in *Salmonella typhimurium. J. of Bacteriology* **169**:2872-2875.

Scholzen, T., and E. Arndt. 1991. Organization and nucleotide sequence of ten ribosomal protein genes from the region equivalent to the spectinomycin operon in the archaebacterium *Halobacterium marismortui*. *Mol. Gen. Genet.* **228**:70-80.

Segall, A., Mahan, M.J., and J.R. Roth. 1988. Rearrangements of the bacterial chromosome: Forbidden inversions. *Science* **241**:1314-1318.

Sensen, C.W., Klenk, H.-P., Singh, R.K., Allard, G., Chan, C.C., Liu, Q.Y., Penny, S.L., Young, F., Schenk, M.E., Gaasterland, T., Doolittle, W.D., Ragan, M.A., and R.L. Charlebois. 1996. Organizational characteristics and information content of an archaeal genome: 156 kb of sequence from *Sulfolobus solfataricus* P2. *Mol. Microbiol.* **22**:175-191.

Sensen, C.W., Charlebois, R.L., Singh, R.K., Klenk, H.-P., Ragan, M.A., and W. F. Doolittle. 1997. Sequencing the genome of *Sulfolobus solfataricus* P2, p. 552-558. In Bruijn, F. J., Lupski, J.R. and G.M. Weinstock (ed.), Bacterial genomes Physical structure and analysis. Chapman & Hall, New York, NY.

Shimmon, L.C., Newton, H., Ramirez, C., Yee, J., Downing, W.L., Louie, A., Matheson, A.T., and P.P. Dennis. 1989. Organization of genes encoding the L11, L1, L10 and L12 equivalent ribosomal proteins in eubacteria, archaebacteria, and eucaryotes. *Can. J. Microbiol.* **35**:164-170.

Siefert, J.L., Martin, K.A., Abdi, F., Widger, W.R., and G.E. Fox. 1997. Conserved gene clusters in bacterial genomes provide further support for the primacy of RNA. *J. Mol. Evol.* **45**:467-472.

Sinden, R.R., and D.E. 1981. Chromosomes in living *Escherichia coli* cells are segregated into domains of supercoiling. *Proc. Natl. Acad. Sci. USA* **78**:224-228.

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., Safer, H., Patwell, D., Prabhakar, S., McDougall, S., Shimer, G., Goyal, A., Pietrokovski, S., Church, G.M., Daniels, C.J., Mao, J., Rice, P., Nölling, J., and J.N. Reeve. 1997. Complete genome *of Methanobacterium thermoautotrophicum* ΔH: Functional analysis and comparative genomics. *J. of Bacteriology.* **179**:7135-7155.

Stettler, R., Erauso, G., and T. Leisinger. 1995. Physical and genetic map of the *Methanobacterium wolfei* genome and its comparison with the updated genomic map of *Methanobacterium thermoautotrophicum* Marburg. *Arch. Microbiol.* **163**:205-210.

St. Jean, A., and R.L. Charlebois. 1996. Comparative genomic analysis of the *Haloferax volcanii* DS2 and *Halobacterium salinarium* GRB contig maps reveals extensive rearrangement. *J. Bacteriology.* **178**:3860-3868.

Tamames, J., Casari, G., Ouzounis, C., and A. Valencia. 1997. Conserved clusters of functionally related genes in two bacterial genomes. *J. Mol. Evol.* **44**:66-73.

Tatusov, R.L., Mushegian, A.R., Bork, P., Brown, N.P., Hayes, W.S., Borodovsky, M., Rudd, K.E., and E.V. Koonin. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli. Current Biology.* **6**:279-291.

Thompson, J.D., Higgins, D.G., and T.J. Gibson. 1994. Clustal W: Improving the sensitivity of progressive multiple alignment through sequence weighting, position specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**:4673-5680.

Tomb, J.F., White, O., Kerlavage, A.R., Clayton, R.A., Sutton, G.G., Fleischmann, R.D., Ketchum, K.A., Klenk, H.P., Gill, S., Dougherty, B.A., Nelson, K., Quackenbush, J., Zhou, L., Kirkness, E.F., Peterson, S., Loftus, B., Richardson, D., Dodson, R., Khalak, H.G., Glodek, A., McKenney, K., Fitzegerald, L.M., Lee, N., Adams, M.D., Hickey, E.K., Berg, D.E., Gocayne, J.D., Utterback, T.R., Peterson, J.D., Kelley, J.M., Cotton, M.D., Weidman, J.M., Fujii, C., Bowman, C., Watthey, L., Wallin, E., Hayes, W.S., Borodovsky, M., Karp, P.D., Smith, H.O., Fraser, C.M. and J.C. Venter. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori. Nature* **388**:539-547.

Trevisanato, S.I., Larsen, N., Segerer, A.H., Stetter, K.O., and R.A. Garrett. 1996. Phylogenetic analysis of the archaeal order of *Sulfolobales* based on sequences of 23S rRNA genes and 16S/23S rDNA spacers. *System. Appl. Microbiol.* **19**:61-65.

Wakagi, T., and T. Oshima. 1985. Membrane-bound ATPase of a thermoacidophilic archaebacterium. *Biochimica et Biophysica Acta.* **817**:33-41.

Weinstock, G.M., and J.R. Lupski. 1997. Chromosomal rearrangements, p. 112-118. In Bruijn, F. J., Lupski, J.R. and G.M. Weinstock (ed.), Bacterial genomes Physical structure and analysis. Chapman & Hall, New York, NY.

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

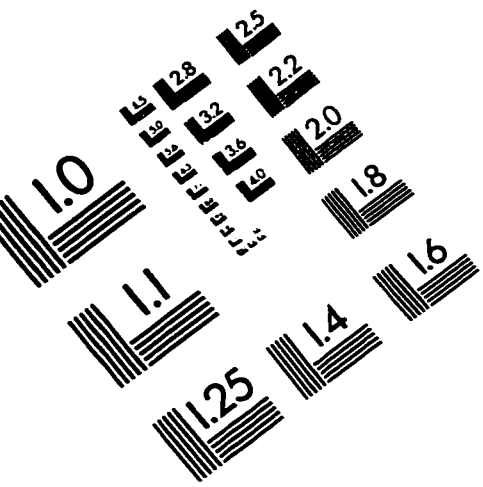
Woese, C.R., Kandler, O. and M.L. Wheelis. 1990. Towards a natural system of organisms: Proposal for the domains Archaea, Bacteria and Eucarya. *Proc. Natl. Acad. Sci. USA.* **87**:4576-4579.

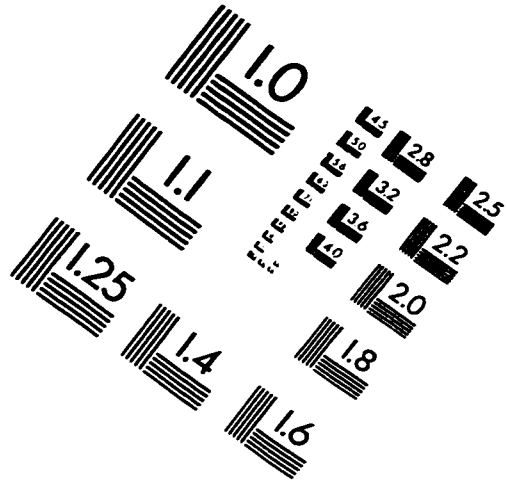
Yamagishi, A., and T. Oshima. 1990. Circular chromosomal DNA in the sulfur-dependent archaebacterium *Sulfolobus acidocaldarius. Nucleic Acids Research.* **18**:1133-1136.

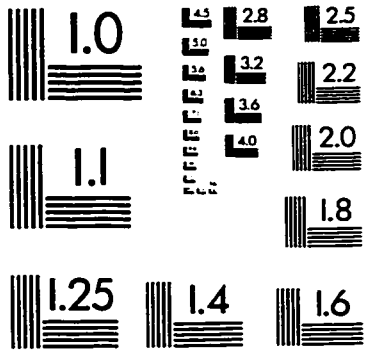
Zillig, W., Stetter, K.O., Wunderl, S., Schulz, W., Priess, H., and I. Scholz. 1980. The *Sulfolobus*-"Caldariella" group: Taxonomy on the basis of the structure of DNA-dependent RNA polymerases. *Arch. Microbiol.* **125**:259-269.

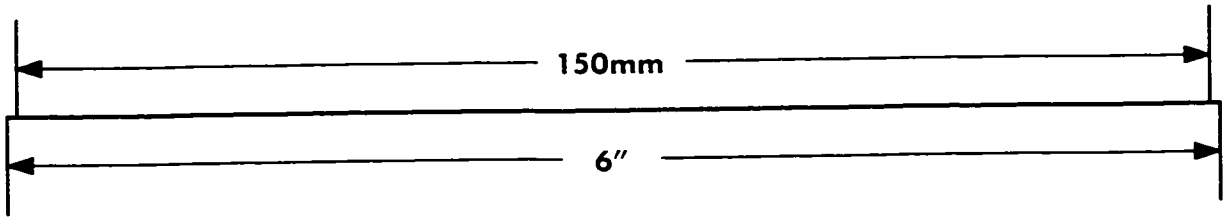
Zuerner, R.L., Herrmann, J.L., and I. Saint Girons. 1993. Comparison of genetic maps for two *Leptospira interrogans* serovars provides evidence for the two chromosomes and intraspecies heterogeneity. *J. Bacteriology.* **175**:5445-5451.

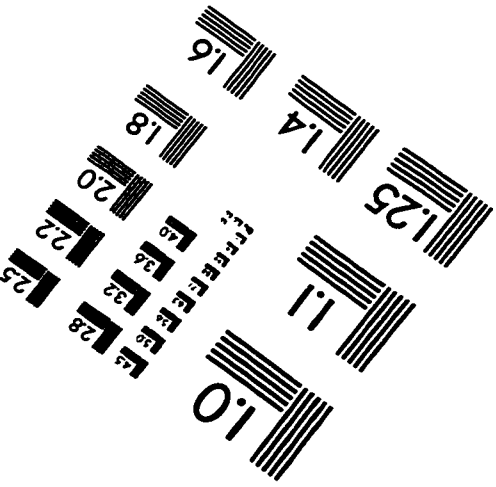
IMAGE EVALUATION
TEST TARGET (QA-3)

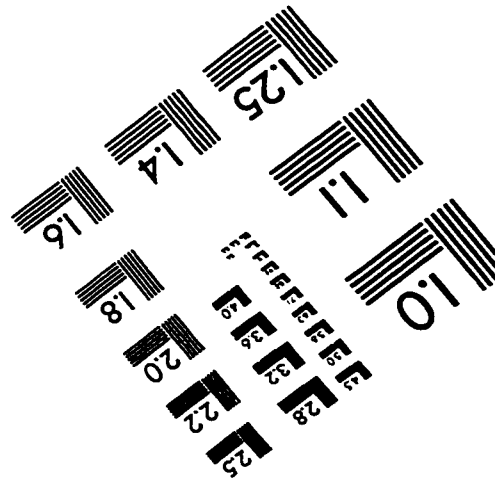












APPLIED IMAGE, Inc
1653 East Main Street
Rochester, NY 14609 USA
Phone: 716/482-0300
Fax: 716/288-5989