FEBS Letters 389 (1996) 88–91 FEBS 17136

Minireview

The *Sulfolobus solfataricus* P2 genome project

Robert L. Charlebois[a,b,*], Terry Gaasterland[a,c,d], Mark A. Ragan[a,e], W. Ford Doolittle[a,f], Christoph W. Sensen[a,e]

[a]*Canadian Institute for Advanced Research, Program in Evolutionary Biology, University of Ottawa, 30 Marie Curie, Ottawa, ON K1N 6N5, Canada*
[b]*Department of Biology, University of Ottawa, 30 Marie Curie, Ottawa, ON K1N 6N5, Canada*
[c]*Mathematics and Computer Science Division, Argonne National Laboratory, 9700 S. Cass Ave., Argonne, IL 60439, USA*
[d]*Department of Computer Science, University of Chicago, 100 E. 58th St., Chicago, IL 60637, USA*
[e]*Institute for Marine Biosciences, National Research Council of Canada, 1411 Oxford Street, Halifax, NS B3H 3Z1, Canada*
[f]*Department of Biochemistry, Sir Charles Tupper Medical Building, Dalhousie University, Halifax, NS B3H 4H7, Canada*

Received 10 May 1996

Abstract Over 800 kbp of the 3-Mbp genome of *Sulfolobus solfataricus* have been sequenced to date. Our approach is to sequence subclones of mapped cosmids, followed by sequencing directly on cosmid templates with custom primers. Using a prototype automated system for genome-scale analysis, known as MAGPIE, along with other tools, we have discovered one open reading frame of at least 100 amino acids per kbp of sequence, and have been able to associate 50% of these with known genes through database searches. An examination of completely sequenced cosmids suggests a clustering of genes by function in the *S. solfataricus* genome.

Key words: Archaea; Genome sequencing; *Sulfolobus solfataricus*; Thermophilic Crenarchaeota

1. Introduction

There is something exotic about an organism that thrives in steaming, sulfurous caldrons, testing the limits of our biochemistry, reminding us of a much harsher distant past. *Sulfolobus solfataricus* is not the most extreme of extremophiles, but it and its genome surely qualify as interesting. This minireview will serve to describe the ongoing Sulfolobus Genome Project and relate its findings to a rapidly growing body of literature on this and related Crenarchaeota.

Interest in the biology of *Sulfolobus* spp. began with the discovery of this unique genus of prokaryotes in the early 1970s [1]. Vague biochemical and morphological similarities with *Thermoplasma* and with the extreme halophiles were noted, the former being another thermoacidophile and the latter also being 'bacteria' lacking peptidoglycan in their cell walls. It was not until the late 1970s though, that sequence data conclusively confirmed the phylogenetic relationship [2,3] and, more importantly, founded the discipline of microbial phylogenetics [4]. Key to our understanding of the evolution of life is the study of *Sulfolobus* and other members of the Archaea. They provide, with the eubacteria and the eukaryotes, three perspectives on biochemistry and molecular genetics. Their specific kinship with eukaryotes [5,6] furthermore promises that the study of Archaea will elucidate much of the basic genetic inventory and physiological capacity from which the eukaryotic cell was able to develop.

Knowledge about *Sulfolobus* spp. and other members of the Sulfolobales has been increasing at a steady rate for the past 15 years. The largest effort, representing roughly 40% of the 350 publications that we have collected, deals with biochemical characterization of proteins or enzymes involved in metabolism, emphasizing respiration, oxidation and oxidative phosphorylation. The study of thermostability is a commonly mentioned goal. Second in terms of publication volume are those concerned with macromolecules, with papers on transcription, translation and DNA metabolism numbering over 100; here the unique features of archaeal gene expression and those features shared with eukaryotes are often emphasized. Roughly 10% of the papers cover microbiological characterization and another 10% describe the cell envelope. The remaining thirty-odd publications are split between industrial applications (desulfurization, bioleaching) and the characterization and development of *Sulfolobus* genetics. Additionally, numerous papers have been published over the years not specifically on *Sulfolobus* spp., but dealing with larger issues of archaeal evolution, or describing and characterizing related systems. Work on *Pyrococcus furiosus*, for instance, is not directly attributable to an interest in *Sulfolobus* spp., but contributes to the global understanding of extreme thermophily and of sulfur-dependent archaeal physiology.

2. Genome project

The Canadian genome project on *S. solfataricus* P2 was not initiated merely to accelerate progress in understanding this species' biology. Certainly, this project will greatly facilitate the hunt for genes and will obviate further sequencing of *S. solfataricus* DNA (NCBI currently lists 40 nucleotide sequences under *S. solfataricus*). It will, by exposing the entire genetic inventory, permit many functions, pathways, processes and regulatory networks to be scrutinized more thoroughly. In essence, the sequence will serve as an important platform from which numerous biochemical, physiological and genetic experiments can be launched. Moreover, *Sulfolobus* is useful for industrial applications, in its capacity to oxidize sulfur (e.g. [7]) and to dissolve pyrite (e.g. [8]), and as a source of a complete set of thermostable enzymes. An increased focus on its biology might therefore not only serve the interests of scientific curiosity in understanding 'life on the edge', but also supply tools of practical value to industry.

*Corresponding author. Fax: (1) (613) 562-5486.
E-mail: robert@bio01.bio.uottawa.ca; Sulfolobus home page: http://www.imb.nrc.ca/imb/sulfolob/sulhom_e.html

PII S0014-5793(96)00525-X

Another reason for choosing *Sulfolobus* was its position in the phylogenetic tree. The dataset of archaeal genes and gene organization obtained from this genome project will greatly enhance our ability to generate and to test hypotheses about the evolution of genes and of genomes. The resolution afforded by the comparison of complete sequences from several genomes may permit us to extrapolate with more confidence towards the common ancestor of prokaryotes and eukaryotes, and even more deeply towards the last common ancestor of all extant life.

Many archaeal genomes are interesting. The specific attraction of *S. solfataricus*, however, was that: (i) considerable investment in *Sulfolobus* spp. research had already been undertaken by others, as outlined above; (ii) it is an aerobic, acidophilic extreme thermophile, somewhat easier to grow in the laboratory than are the anaerobes; (iii) it can grow chemolithotrophically and autotrophically as well as heterotrophically [1], thus elaborating alternative energy-harvesting strategies and complete biosynthetic capability; (iv) development of genetic tools (e.g. [9,10]), which will be necessary for downstream genetic manipulations, is underway; (v) its chromosome is among the largest in the Archaea ([11] and unpublished), thus the risk of this genome having dispensed with some of the ancestral archaeal genes is reduced; (vi) such a large chromosome may furthermore contain genes in addition to the core set, and include secondary or peripheral metabolism potentially useful in biotechnology; (vii) its genome has a low G+C content (mean of 37%) and thus is easier to sequence than, say, that of the extreme halophiles whose well-studied genomes [12,13] are G+C-rich; and (viii) open reading frames (ORFs) discovered by sequencing are more likely to be real genes in a low G+C context, since given the higher chance probability of randomly generated stop codons, long non-coding (statistical) ORFs occur less often.

Our approach to sequencing the genome of *S. solfataricus* P2 [14] is shown in Fig. 1. In brief, mapped cosmid subclones of genomic DNA are further subcloned as randomly sheared fragments into pUC18 for sequencing using fluorescent automated techniques. Following the initial shotgun sequencing phase (which produces a redundancy between 2 and 3), we apply custom primers to cosmid templates for direct cosmid sequencing to link contigs, fill single-stranded gaps and resolve ambiguities. We discriminate between four distinct phases in the sequencing of a cosmid: (i) the primary sequencing phase (many gaps, many ambiguities); (ii) the linking phase (a few gaps, fewer ambiguities); (iii) the polishing phase (one contig, approximately one ambiguity per kbp); (iv) and the finished phase (publication-quality, < 1 error per 5 kbp). The distribution of the sequencing effort among the four different phases is easily adjustable according to need (Fig. 2). It is now clear that the entire *S. solfataricus* genome cannot be cloned as cosmids, but a number of alternative subdivisions of the genome are available, such as macrorestriction fragments or phage-based clones.

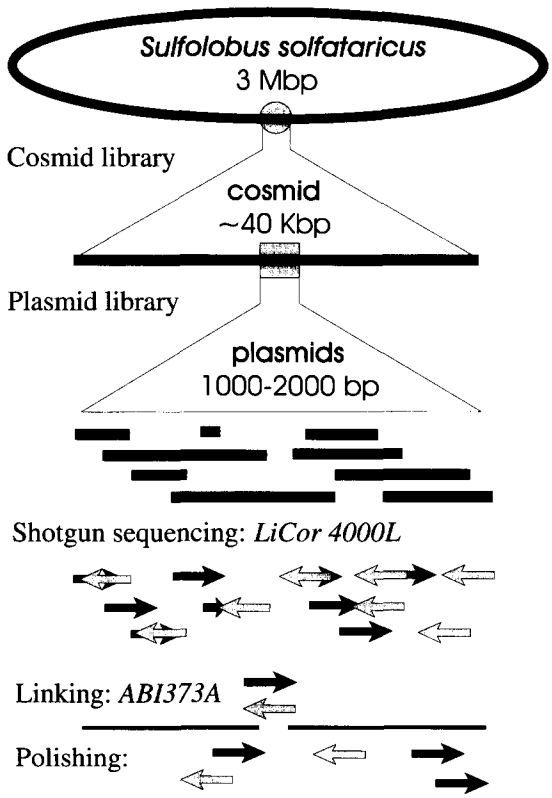

Fig. 1. Sequencing strategy employed within the Sulfolobus Genome Project.

3. Current database

At the time of writing this paper, in April of 1996, we had sequenced over 800 kbp of the 3 Mbp *S. solfataricus* genome (Fig. 2). Since the sequences are organized in discrete cosmid-sized or larger contig-sized strings, we have been able to identify genes since the beginning of the project. Modest financial resources limit us to a current production rate of 80–100 kbp of new sequence per month, yet unlike total genomic shotgun projects [15,16], this sequence assembles into ORFs continually rather than all at once near the end of the project. Our present goal is to produce a linked 3-Mbp circular sequence by the end of 1997, generating a complete nucleotide-resolution genetic map from which individual genes of interest can easily be extracted for further study.

Bioinformatics is the means by which these endless strings of nucleotides are made meaningful. In a joint effort with the *Sulfolobus* team, an automated system for comprehensive genome-scale sequence analysis, called MAGPIE [17], is being developed at Argonne National Laboratory. MAGPIE distributes requests to a variety of local and remote tool servers, assimilates the responses, and generates reports that are queryable and browsable across the World Wide Web. This system efficiently and automatically updates analyses of our local *Sulfolobus* database (as well as other local genome databases) with respect to the public databases, as each accretes new information. The findings become available immediately to our distributed collaborative network and eventually to the entire research community.

Of the more than 800 kbp sequenced thus far, 448 kbp are in the form of completely linked contigs. These contain nine known RNA genes and 491 ORFs of at least 100 amino acids in size. Since the G+C content of the genome is low, few 'statistical ORFs' are expected to occur, and indeed among these 491 ORFs only 61 (12%) are conflicting (i.e. derive from the same sequence in a different reading frame). It is presently difficult to estimate how many of the remaining non-conflicting ORFs encode genes, since this depends on how much non-

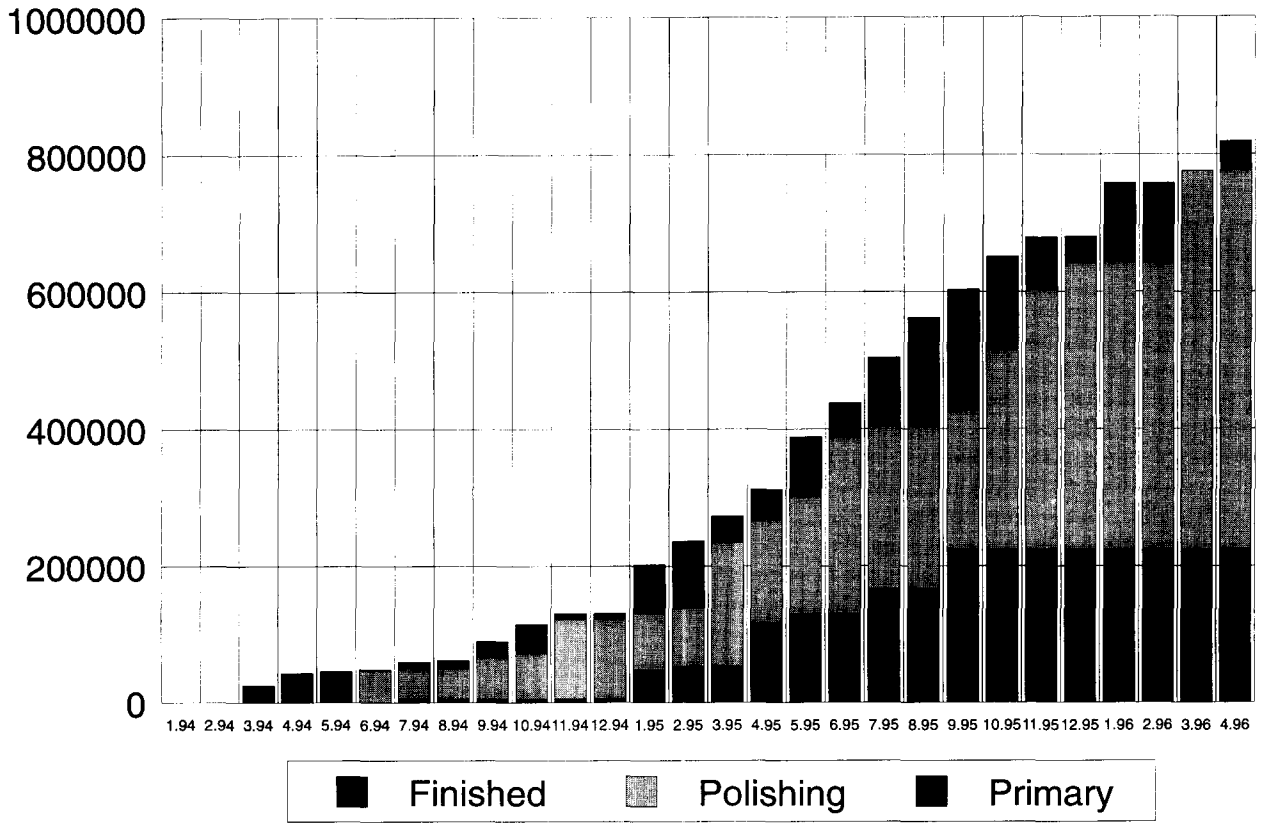

Fig. 2. Sequencing progress. The cumulative number of base pairs sequenced is displayed month-by-month since the initiation of sequencing in early 1994. The proportion of sequences in the three principal phases of completion are shown; we recognize a further subdivision – 'linking' – within the polishing phase (see text) though both activities are strategically similar (Fig. 1). Very recent acquisition of additional automated sequencing equipment has pushed our present productivity to 80–100 kbp per month for April 1996 and beyond.

coding DNA is present in this genome. Codon usage tables and other measures of likelihood are under development to address this issue. We know at least that gene density is between 0.5 and 1.0 per kbp, since ORFs are packed at an average of one per kbp, and we succeed in identifying one gene per 2 kbp through database similarity (Table 1). Most conflicting ORFs tend to be small, yet our unidentified ORFs range in size up to the largest ORF discovered in the genome so far (1068 codons).

4. Clustering of genes

There appears to be clustering of genes by function in the *S. solfataricus* genome (Table 1). Upon classifying the linked contigs' 221 database-identified genes by function according to Riley's system [18], contig-specific themes become apparent. Contigs c04–c05 and c10, for instance, are predominantly concerned with macromolecules, whereas contig c08–c09 deals with the biosynthesis of small molecules and contigs c14 and

Table 1
Classification of *S. solfataricus* ORFs matching database entries

Cosmids in linked contigs →	c06 c02 c01	c04 c05	c08 c09	c10	c13	c14	c18	c19	c97	Totals
Size of linked contig in kbp →	100	56	58	40	34	44	35	39	42	448
I. Intermediary metabolism	13	2	6	1	0	21	8	8	12	71
II. Biosynthesis of small molecules	2	3	29	3	1	2	0	5	3	48
III. Macromolecule metabolism	2	26	5	24	0	3	1	2	1	64
IV. Cell structure	0	1	0	0	0	0	0	0	0	1
V. Cellular processes	6	4	0	2	1	2	5	4	0	24
VI. Other functions	1	0	0	0	0	0	0	0	0	1
Matching unidentified database ORFs	4	1	3	0	0	1	0	0	3	12
Totals	28	37	43	30	2	29	14	19	19	221
Strong database matches per kbp	0.28	0.66	0.74	0.75	0.06	0.66	0.40	0.49	0.45	0.493

c97 are primarily catabolic. Much but not all of this functional clustering can be attributed to putative operons. Summing the number of genes found in each category gives us a first glimpse at how *S. solfataricus* invests in each type of activity. Even from our limited sample of archaeal DNA, there is close similarity in the relative usage of DNA for catabolism, biosynthesis, macromolecules, and cellular processes compared to *Escherichia coli* [18] and *Haemophilus influenzae* [16]. Very few of the genes in our *S. solfataricus* sample match database functions relating to cell structure or to 'other functions'. These categories are more idiosyncratic, especially with regards to the cell envelope.

Another kind of uneven distribution of genes is evident in the genome when one examines the density of database matches within individual contigs (Table 1). Contigs with functional clustering tend to have a series of matches with database entries, consistent with the bias that if one gene in a pathway or in a complex is characterized, it is likely that others are as well. However, poor scores in some of the contigs, especially c13 and c06–c02–c01, are not simply due to specific deficiencies in the data banks. These contigs are considerably enriched in repeated sequences – especially insertion sequences – which in c13 consume at least 50% of the DNA. Contig c06–c02–c01 contains at least nine insertion sequences or IS-like elements and several shorter repeated elements. Contigs c18 and c97, with their slightly below-average score for database matches, possess five and three major repeats, respectively. The other contigs listed in Table 1 have no known repeats.

It is interesting and relevant to note that only 12 of the 221 database matches were against hypothetical proteins, despite the ability to search against extensive tracts of genomic sequence. Many more of the ORFs from *H. influenzae* and *Mycoplasma genitalium* matched hypothetical proteins from other organisms (26 and 15% of matches, respectively) [16], since these genomes are phylogenetically close to model organisms whose genomic sequences are well sampled. In contrast, *S. solfataricus*, an archaeon, currently has no such extent of sequence data with which those of its genes that are less-well conserved can match. *Synechocystis* sp. PCC6803, at an intermediate distance from the major model bacterial genomes, has 10% of ORFs matching hypothetical proteins [19]. It would seem that most of the highly conserved and ubiquitous gene families have already been identified, as has been previously suggested [20]. Many ORFs remain which are unidentified because they are specific to a lineage, thanks either to recent recruitment or to a loss of detectable sequence similarity through time. Hundreds of ORFs scoring against hypothetical proteins is the situation that we can expect once more pairs of related genomes are sequenced. There is already an indication that the Archaea share numerous genes specific to their lineage [21]. This large collection of unassigned genes will demand direct genetic and biochemical characterization, and the establishment of a thermophilic archaeal model organism with a status equivalent to *E. coli* or to yeast. A collective effort from various strategic phylogenetic points can eventually drive the global number of unassigned ORFs towards the level of background noise.

5. Genome sequencing

Genomic sequencing, the new paradigm in biology, serves as a valuable resource to be embraced and exploited by the research community. Throughout the phylogenetic tree of life, long strings of nucleotides and deduced amino acids are being elucidated, and this information will effectively revolutionize molecular biology. We will be able to perform many experiments in the computer to supplement those that we do at the bench in our quest to understand biology. Genomics is currently about gathering base pairs and is thus primarily a technical feat. The future of genomics, however, is to become a new science for studying integrated systems of information.

Acknowledgements: The work described in this paper is funded mainly by the Canadian Genome Analysis and Technology Program, the National Research Council of Canada, the Canadian Institute for Advanced Research and the Medical Research Council of Canada. T.G. is funded by the U.S. Department of Energy under Contract W-31-103-Eng-38. We gratefully acknowledge the efforts of our technical team. Issued as NRCC number 39716.

References

[1] Brock, T.D., Brock, K.M., Belly, R.T. and Weiss, R.L. (1972) Arch. Mikrobiol. 84, 54–68.
[2] Woese, C.R., Magrum, L.J. and Fox, G.E. (1978) J. Mol. Evol. 11, 245–252.
[3] Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A., Wolfe, R.S., Balch, W.E., Tanner, R.S., Magrum, L.J., Zablen, L.B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B.J., Stahl, D.A., Luehrsen, K.R., Chen, K.N. and Woese, C.R. (1980) Science 209, 457–463.
[4] Woese, C.R. (1987) Microbiol. Rev. 51, 221–271.
[5] Brown, J.R. and Doolittle, W.F. (1995) Proc. Natl. Acad. Sci. USA 92, 2441–2445.
[6] Doolittle, R.F., Feng, D.-F., Tsang, S., Cho, G. and Little, E. (1996) Science 271, 470–477.
[7] Kargi, F. and Robinson, J.M. (1982) Appl. Environ. Microbiol. 44, 878–883.
[8] Lindström, E.B., Wold, S., Kettaneh-Wold, N. and Sääf, S. (1993) Appl. Microbiol. Biotechnol. 38, 702–707.
[9] Schleper, C., Holz, I., Janekovic, D., Murphy, J. and Zillig, W. (1995) J. Bacteriol. 177, 4417–4426.
[10] Keeling, P.J., Klenk, H.-P., Singh, R.K., Feeley, O., Schleper, C., Zillig, W., Doolittle, W.F. and Sensen, C.W. (1996) Plasmid 35, 141–144.
[11] Kondo, S., Yamagishi, A. and Oshima, T. (1993) J. Bacteriol. 175, 1532–1536.
[12] Cohen, A., Lam, W.L., Charlebois, R.L., Doolittle, W.F. and Schalkwyk, L.C. (1992) Proc. Natl. Acad. Sci. USA 89, 1602–1606.
[13] St. Jean, A., Trieselmann, B.A. and Charlebois, R.L. (1994) Nucleic Acids Res. 22, 1476–1483.
[14] Sensen, C.W., Charlebois, R.L., Singh, R.K., Klenk, H.-P., Ragan, M.A. and Doolittle, W.F. (1996) in: Bacterial Genomes: Physical Structure and Analysis (De Bruijn, F.J., Lupski, J.R. and Weinstock, G. eds.) Chapman and Hall, New York, in press.
[15] Fleischmann, R.D. et al. (1995) Science 269, 496–512.
[16] Fraser, C.M. et al. (1995) Science 270, 397–403.
[17] Gaasterland, T. and Sensen, C.W. (1996) Trends Genet. 12, 76–78.
[18] Riley, M. (1993) Microbiol. Rev. 57, 862–952.
[19] Kaneko, T., Tanaka, A., Sato, S., Kotani, H., Sazuka, T., Miyajima, N., Sugiura, M. and Tabata, S. (1995) DNA Res. 2, 153–166.
[20] Green, P., Lipman, D., Hillier, L., Waterston, R., States, D. and Claverie, J.-M. (1993) Science 259, 1711–1716.
[21] Ouzounis, C., Kyrpides, N. and Sander, C. (1995) Nucleic Acids Res. 23, 565–570.