Extremophiles (2004) 8:291–299
DOI 10.1007/s00792-004-0388-1

ORIGINAL PAPER

Xiu-Feng Wan · Susan M. Bridges · John A. Boyle

Revealing gene transcription and translation initiation patterns in archaea, using an interactive clustering model

Received: 16 May 2003 / Accepted: 11 March 2004 / Published online: 19 May 2004

Abstract An interactive clustering model based on positional weight matrices is described and results obtained using the model to analyze gene regulation patterns in archaea are presented. The 5′ flanking sequences of ORFs identified in four archaea, *Sulfolobus solfataricus*, *Pyrobaculum aerophilum*, *Halobacterium* sp. NRC-1, and *Pyrococcus abyssi*, were clustered using the model. Three regular patterns of clusters were identified for most ORFs. One showed genes with only a ribosome-binding site; another showed genes with a transcriptional regulatory region located at a constant location with respect to the start codon. A third pattern combined the previous two. Both *P. aerophilum* and *Halobacterium* sp. NRC-1 exhibited clusters of genes that lacked any regular pattern. *Halobacterium* sp. NRC-1 also presented regular features not seen in the other organisms. This group of archaea seems to use a combination of eubacterial and eukaryotic regulatory features as well as some unique to individual species. Our results suggest that interactive clustering may be used to examine the divergence of the gene regulatory machinery in archaea and to identify the presence of archaea-specific gene regulation patterns.

Communicated by K. Horikoshi and F. Robb

X.-F. Wan · S. M. Bridges (✉)
Department of Computer Science and Engineering,
Mississippi State University,
Mississippi State, MS 39762, USA
E-mail: bridges@cse.msstate.edu
Tel.: +1-662-3257505
Fax: +1-662-3258997

J. A. Boyle
Department of Biochemistry and Molecular Biology,
Mississippi State University,
Mississippi State, MS 39762, USA

Present address: X.-F. Wan
Digital Biology Laboratory,
University of Missouri Columbia,
Columbia, MO 65211, USA

Keywords Archaea · Clustering · *K*-means · Transcription initiation · Translation initiation

Introduction

Archaea constitute one of the three major domains of life on earth. They are a widespread group of organisms found in extreme environments such as the deep ocean, bogs, salt brines, and hot acid springs, and also in subsurface marine waters (Delong et al. 1994). Pelagic archaea were reported to constitute more than 30% of the prokaryotic biomass in coastal Antarctic surface waters. More than 30 species of archaea have been isolated from a single pond in Yellowstone National Park (Barns et al. 1996). Archaea may be divided into three subgroups: Euryarchaeota, Crenarchaeota, and Korarchaeota. As of 1 December 2003, the National Center for Biotechnology Information (NCBI) lists 17 complete archaeal genomes that have been sequenced, all belonging to Euryarchaeota or Crenarchaeota (Barns et al. 1996). Two Euryarchaeota, *Pyrococcus abyssi* and *Pyrobaculum aerophilum*, and two Crenarchaeota, *Halobacterium* sp. NRC-1 and *Sulfolobus solfataricus*, were selected for this study (Table 1).

Transcription initiation patterns in archaea seem to be related more closely to those of eukaryotes than to eubacteria (Soppa 1999a, b). However, three groups of transcription-associated proteins have been identified in archaea: one group having homology with prokaryotes, another group having homology with eukaryotes, and the third group having homology with both prokaryotes and eukaryotes. Recently, several homologues of bacterial transcriptional factors, such as MDR1 (metal-dependent repressor) (Bell et al. 1999), leucine-responsive regulatory protein (Lrp) (Kyrpides and Ouzounis 1995; Dahlke and Thomm 2002), a heat shock regulator (Phr) (Vierke et al. 2003), and transcriptional regulator of mal operon (Trm) (Lee et al. 2003) have been identified and characterized in archaea. More recently, Ouhammouch

Table 1 Archaeal genomic data analyzed

Genome name	Gene number	Order	GenBank accession No.	References
Sulfolobus solfataricus	2,977	Crenarchaeota	NC-002754	She et al. 2001
Pyrococcus abyssi	1,765	Euryarchaeota	NC-000868	Natale et al. 2000
Pyrobaculum aerophilum	2,605	Euryarchaeota	NC-003364	Fitz-Gibbon et al. 2002
Halobacterium sp. NRC-1	2,058	Crenarchaeota	NC-002607	Ng et al. 2000

et al. (2003) have demonstrated that archaea possess a eukaryote-like positive regulation transcription apparatus consisting of a cognate bacterial-type regulator that facilitates recruitment of TATA-binding protein as a mechanism of transcriptional activation. Thus, archaeal transcription systems may exhibit their own special transcription patterns in addition to having patterns similar to eubacteria and eukaryotes (Kyrpides and Ouzounis 1999).

Two types of translation initiation mechanisms have been reported in archaea. Leadered translation—similar to the translation process in eubacteria—has been shown to occur for internal genes of operons in archaea (Tolstrup et al. 2000; Slupska et al. 2001). This type of translation initiation involves the binding of 16S rRNA to the Shine–Dalgarno sequence of the mRNA. The genes in this group have a G-rich region in their 5′ flanking region, which corresponds to the Shine–Dalgarno consensus sequence. Archaeal translation can also employ leaderless translation initiation similar to eukaryotes. Leaderless translation initiation involves scanning for the first start codon along the transcript (Kozak 1999). The first gene in operons and isolated genes in archaea were reported to have this eukaryotic-like translation system (Tolstrup et al. 2000).

Boyle and Boyle (2003) reported different upstream patterns for different archaeal genomes. In *S. solfataricus* and *P. abyssi*, there is a G-rich region centered at position −10 (counting from the translation start codon). In addition, there is an A-rich region centered at −34, followed by a T-rich region around −29, followed by another A-rich region around −25. *Halobacterium* sp. NRC-1 and *P. aerophilum* reveal different upstream patterns, but neither has a G-rich region.

Our goal is to identify translation and transcription initiation patterns in archaea by clustering regulatory sites of genes for entire genomes, using a newly developed interactive clustering model (Wan et al. 2002) based on the positional weight matrix (PWM) (Staden 1984) representation of gene sequences. The interactive clustering method facilitates the identification of qualitative differences in the signals in complete genomes, and the quantitative results (size of the resulting clusters) provide a rough estimate of the abundance of certain types of signals. This method can provide guidance and support for hypothesis generation regarding the existence of transcription and translation signals in genomes.

We have identified groups of genes with clearly different patterns in regulatory regions. Our results suggest archaea may have gene-regulation patterns other than those typical of eubacteria or eukaryotes.

Materials and methods

Genomes

Genomic sequences of four archaea (Table 1) and their associated protein annotation tables were downloaded from GenBank, NCBI (http://www.ncbi.nlm.nih.gov). The upstream sequence of each ORF was extracted using a Perl program.

PWM and feature-vector representation

The PWM approach has been widely applied since its introduction to sequence analysis by Staden (1984). PWM methods have proven to be very promising in sequence analysis (Bucher 1990; Holberton and Marshall 1995; Levy et al. 2001; Liu et al. 2001). In the work reported in this paper, a PWM was used to represent the probability of each nucleotide (A, T, G, C) at each position in a set of sequences. Given a set of sequences $S=\{S_1, S_2, S_3, \ldots, S_m\}$, the PWM entry for a base b at position p can be computed as

$$M_{b,p} = \log\left(f_{b,p}/f_b\right) \quad (1)$$

where $f_{b,p}$ is the frequency of base b at position p and f_b is the expected frequency of the base. The feature vector for a particular sequence is constructed by selecting the probability of occurrence of the base in each position in the sequence from the PWM for each position. A window size is specified to encompass the region of interest to the scientist.

Interactive clustering model

The k-means and its many variants are widely used clustering algorithms (Jain et al. 1999; Han and Kamber 2000). Given objects $O=\{o_1, o_2, \ldots, o_n\}$, the k-means clustering algorithm partitions them into k-clusters based on the distance D_{ij} between object o_i and o_j, $o \in O$, $1 \le i \le n$, $1 \le j \le n$. The distance of the cluster members to the cluster center is computed by

$$\overline{D} = \frac{1}{m}\sum_{p=1}^{m}\sqrt{\sum_{q=1}^{d}\left(x_{p,q}-x_{c,q}\right)^2} \quad (2)$$

where m is the number of objects in the cluster, d is the dimension of the feature vectors, and x_c is the cluster center. The center of a cluster C_u is updated by

$$C_{u,q} = \frac{1}{m}\sum_{p=1}^{m} x_{p,q} \quad (3)$$

where x_p is an object in cluster C_u, $1 \le q \le d$, and d is the dimension of the feature vectors. A shortcoming of the k-means algorithm is that one must specify or empirically determine the number of clusters. We overcome this problem by applying k-means iteratively, with $k = 2$ at each step. Used in conjunction with a visualization tool, our approach allows the scientist to identify clusters of regulatory patterns by interacting with the clustering process.

The interactive clustering model (Wan et al. 2002) combines PWMs and k-means clustering in an iterative fashion (Fig. 1). At each step, the sequence data are clustered, with $k = 2$ based on a specified window size. A visualization of the PWM of each cluster is then generated for inspection (see Fig. 2 for examples of the visualization). Clusters exhibiting homogeneous patterns (e.g., only a G-rich region or a TATA box) are not clustered further. If a cluster is not clearly homogeneous, it is reclustered. If no new patterns emerge when clustering is reapplied, the new clusters are discarded. The result of this process is a binary tree (see Fig. 4 for an example). Prior to reclustering, users can also change the window size if they wish to focus the reclustering on a specific area of the sequence. In our experiments, we have begun clustering with a window range of −48 to −1 and sometimes reduce the range to −48 to −25 or −24 to −1 during the clustering process. The website for this clustering model is http://www.cse.msstate.edu/~bridges/CLU/Cluster/cluster.html. The model is not fully automated at this time.

Process for analysis of clustering results

A G-rich region centered at position −10 that represents the Shine–Dalgarno consensus is referred to as a "G box" in the remainder of this paper. Similarly, a pattern of As, Ts, then As related to the TATA promoter sequence is referred to as an "A box." A cluster with only one or both G and A boxes and no other apparent pattern is treated as a single cluster. A cluster with neither an A nor G box is also treated as a cluster if further clustering with different windows sizes does not reveal additional patterns.

For one experiment, we separated the genes of *S. solfataricus* into two groups: putative genes and confirmed genes. Putative genes are those genes designated as "putative," "hypothetical," or "conserved hypothetical" in the protein annotation table. We found the patterns identified in both groups were very similar. In subsequent experiments, we have assumed that all annotated ORFs in the genomes are genes.

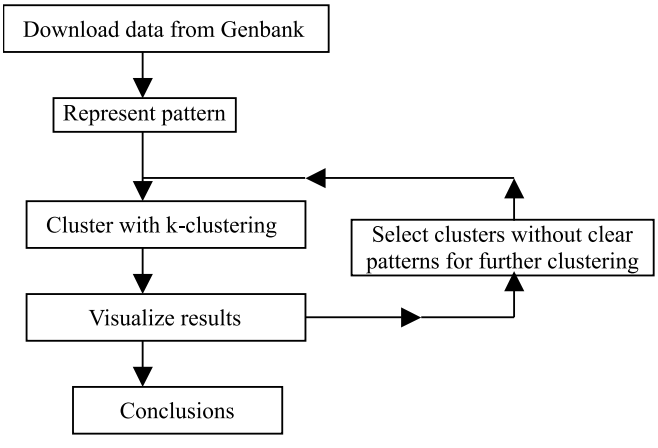


Fig. 1 The interactive clustering model

Results

Histograms of the distances between genomes

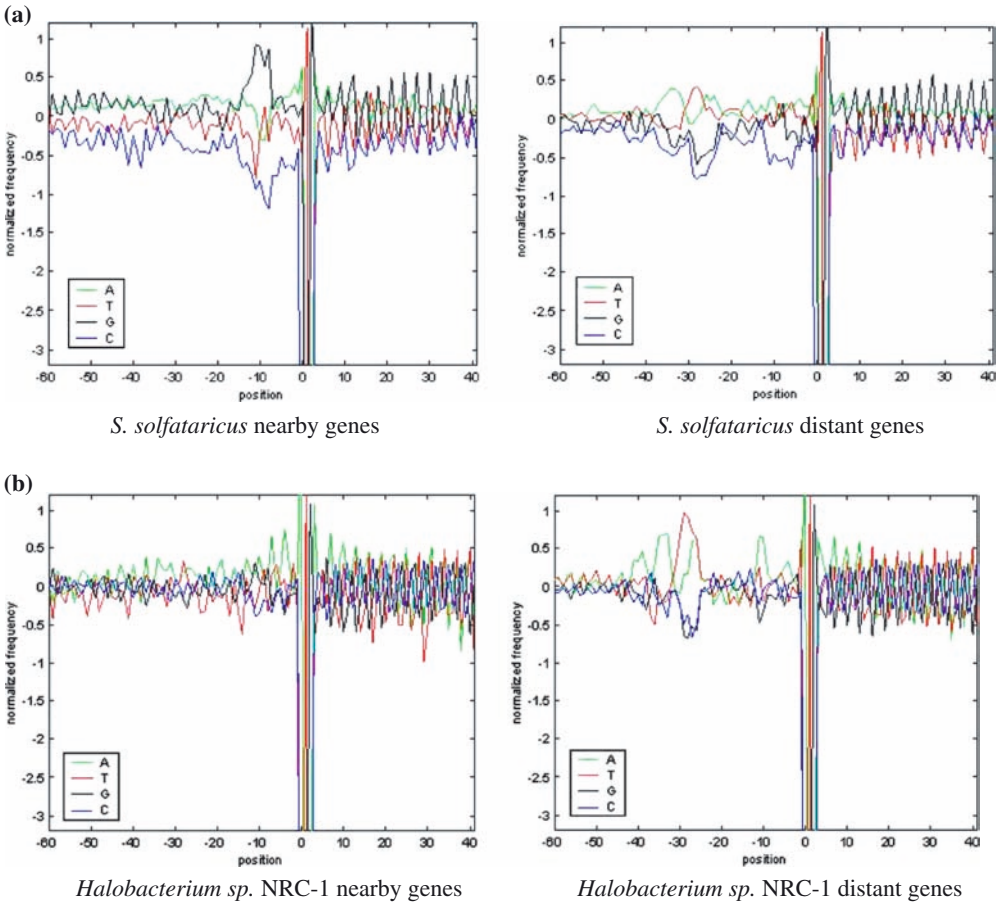
A histogram of the gene separation distances in each genome was plotted (Fig. 3), and the observed distribution was used to group the genes into the classes "distant" and "nearby." Distant genes were defined to be those whose start codon was separated from the termination codon of the nearest upstream gene by more than 25 bp or less than −25 bp (in cases where the genes overlap). All other genes were labeled nearby. The PWMs of the distant and nearby genes for two archaeal species are shown in Fig. 2. In *Sulfolobus solfataricus*, the nearby genes have a G box centered at position −10, and the distant genes have an A box centered at position −30 (Fig. 2a). In *Halobacterium* sp. NRC-1, the nearby genes have a weak A box around position −30, and the distant genes have a strong A box around position −30 and an unusual signal around −10 (Fig. 2b).

Clustering results

All of the clustering results described in this section will be available at http://www.cse.msstate.edu/~bridges/CLU/ArchaeaResults.html, including a list of all genes assigned to each cluster. In the following discussion, we present the binary tree that resulted from our clustering process for each genome and visualizations of the PWMs corresponding to each cluster (node in the tree). In each figure, the number of sequences that each node represents is given in parentheses. The tree nodes are numbered for identification purposes.

Sulfolobus solfataricus

The PWM of all genes of *S. solfataricus* contains a G box centered at −10 and an A box centered at −30 (node 1 in Fig. 4). More detailed results are available at http://



S. solfataricus nearby genes | *S. solfataricus* distant genes

Halobacterium sp. NRC-1 nearby genes | *Halobacterium sp.* NRC-1 distant genes

Fig. 2a, b Example positional weight matrices of nearby genes and distant genes in *Sulfolobus solfataricus* (**a**) and *Halobacterium* sp. NRC-1 (**b**). A distant gene is an ORF that has a distance over 25 bp or less than −25 bp from the previous gene. Otherwise, the gene is classified as a nearby gene

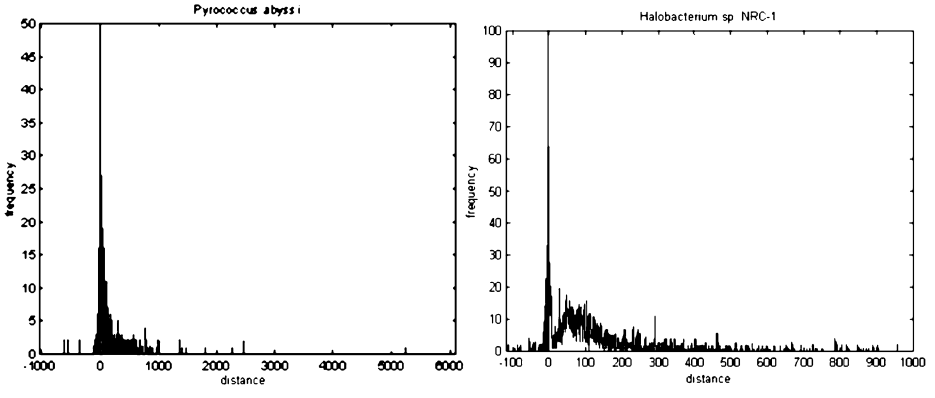


Fig. 3 Histogram of distances between ORFs in *Pyrococcus abyssi* and *Halobacterium* sp. NRC-1. The distance is the gap between the translation start codon of the ORF and the translation stop codon of its neighborhood upstream ORF. A negative value indicates that two ORFs overlap

www.cse.msstate.edu/~bridges/CLU/SUL/sul.html. The initial binary split of this cluster, using a window of −48 to −1, yields two groups (nodes 2 and 3) with quite different patterns. The first cluster (node 2) has an A box and the second cluster (node 3) has a G box (Fig. 4). These two groups were analyzed for distribution of nearby genes and distant genes. Division of the cluster containing the A box (node 2) on the basis of proximity yields two distinct clusters (nodes 4 and 5). The nearby cluster (node 4) exhibits both an A and a weak G box, while the distant cluster (node 5) shows only an A box. Further clustering of the nearby genes, using a window of −24 to −1 yields two clusters—one (node 8) with 146, revealing only an A box, and the other with 142 genes showing both an A and a G box. These could not be further clustered into subclusters with different patterns. Division of the node 3 cluster into nearby and distant genes gave two groups with G boxes only (Fig. 4).

Pyrococcus abyssi

The PWM of *Pyrococcus abyssi* (Fig. 5) has what appears to be superimposed A boxes, one centered at −40 and the other at −30. Initial clustering shows one

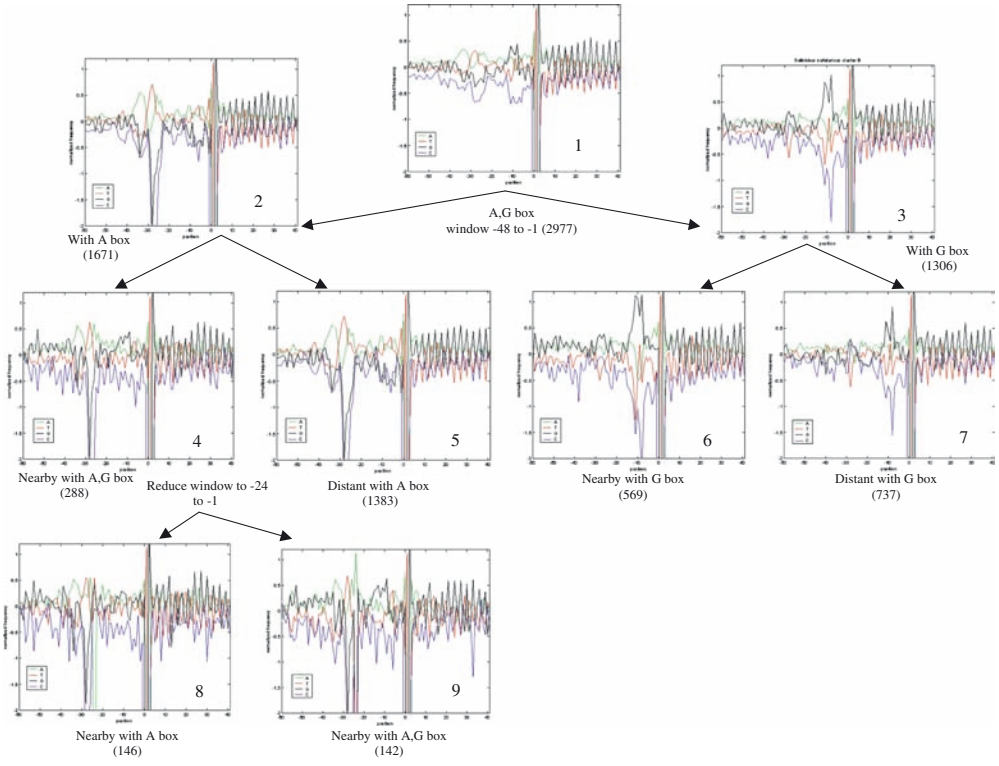


Fig. 4 Clustering results for *S. solfataricus*

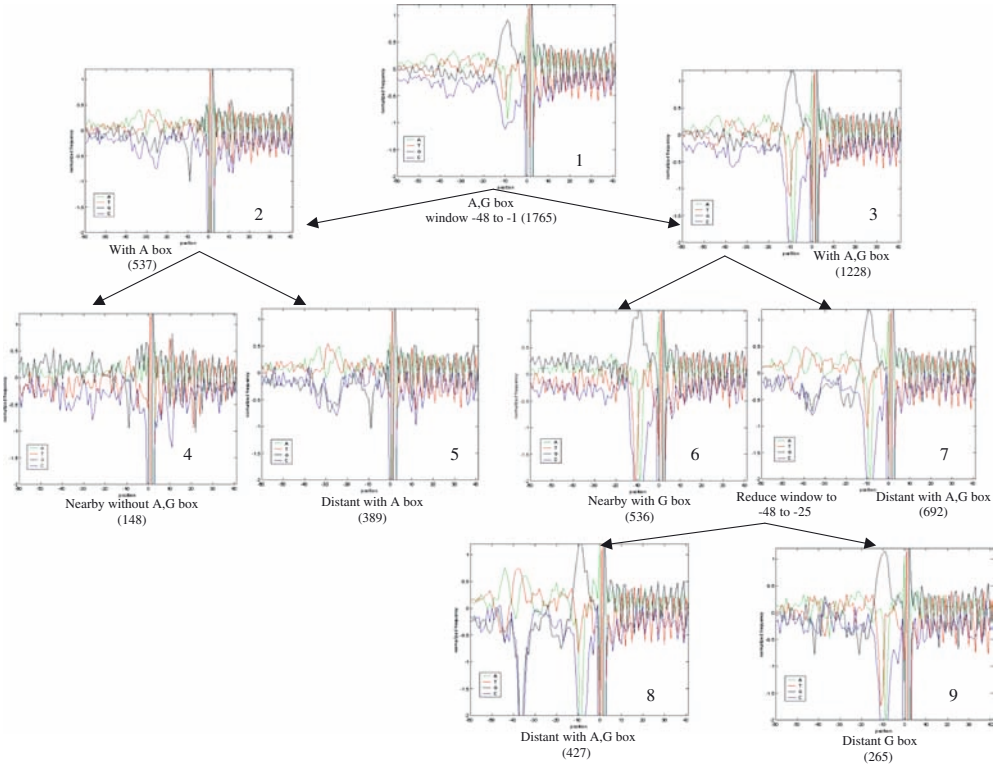


Fig. 5 Clustering results for *P. abyssi*

group with only an A box centered at −30 (node 2). Separation of this node into nearby and distant genes shows that the nearby genes do not retain any obvious pattern, while the distant genes maintain the A box. The other cluster, containing the majority of genes (1,228 genes), looks similar to the initial pattern of all genes. Separation into nearby and distant genes provided us further pattern information. Nearby genes have only the G box at −10 (node 6). The distant genes have both superimposed A boxes and the G box at −10 (node 7).

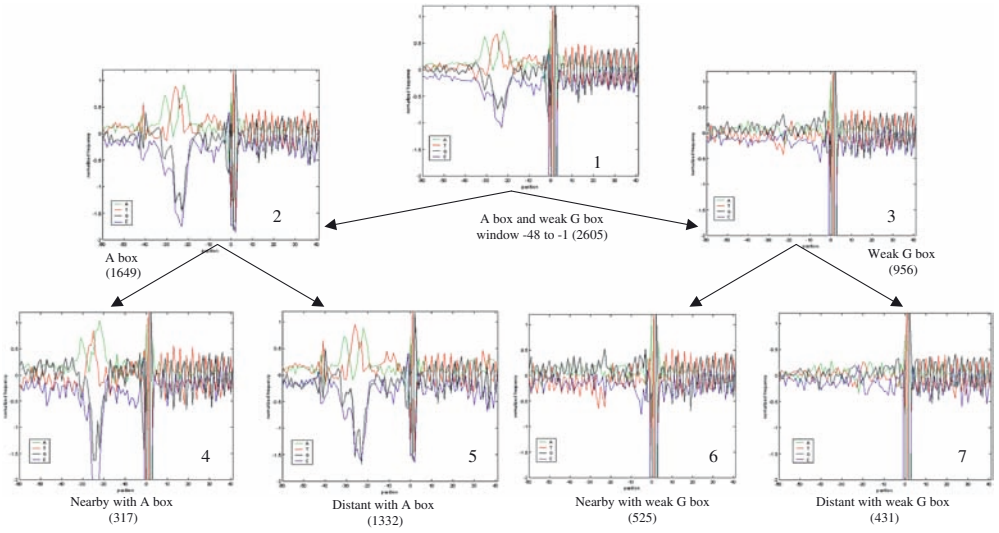


Fig. 6 Clustering results for *Pyrobaculum aerophilum*

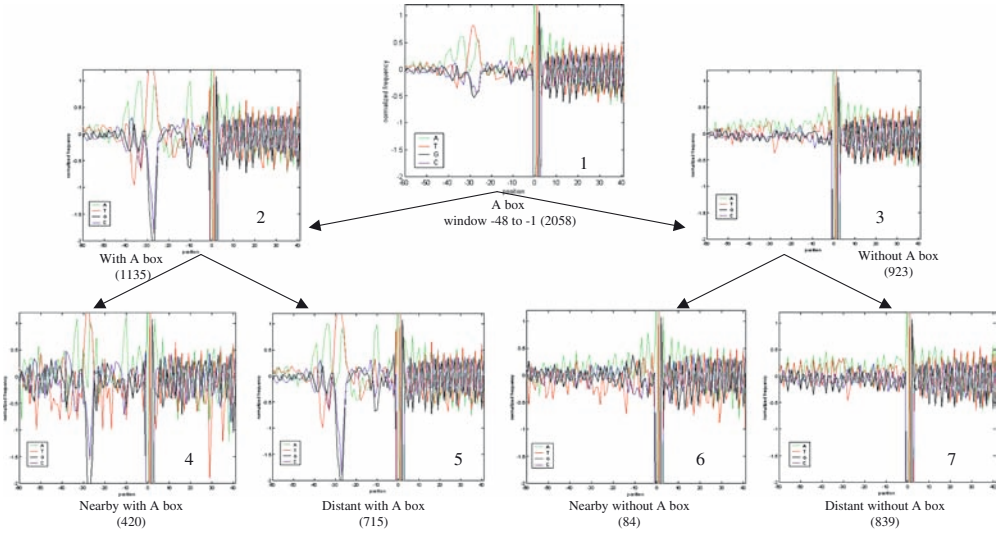


Fig. 7 Clustering results for *Halobacterium* sp. NRC-1

Clustering of node 7 with a window of −48 to −25 gives us 265 genes (node 9) with only a G box and 427 genes with the same overall pattern as node 7.

Pyrobaculum aerophilum

The PWM of *Pyrobaculum aerophilum* (Fig. 6) shows an A box centered at −27. Initial clustering produced one group of genes with an A box at −27 and a second group of genes with a weak G box at −4. Subdivision of these groups by relative position provided little further information (nodes 4–7, http://www.cse.msstate.edu/~bridges/CLU/PYR/pyr.html; Fig. 6).

Halobacterium sp. NRC-1

The genome of *Halobacterium* sp. NRC-1 (Fig. 7) has an upstream pattern that differs from those found in *S. solfataricus*, *P. abyssi*, and *P. aerophilum*. It has a typical A box centered at −30 and three A-rich regions centered at positions −4, −10, and −40. The clustering generated two groups of genes. One group of 1,135 genes (node 2) maintains all features of the total genomic ORFs except for the region at −4. The other group (node 3) has only this −4 A region. Nearby and distant classification was not informative for the node 3 genes. Using this separation, node 2 ORFs split into those with the −40 A peak and those without it. Both nodes 4 and 5 retained the A box at −30 and the A peak at −10. Further clustering was not useful (http://www.cse.msstate.edu/~bridges/CLU/HAL/hal.html; Fig. 7).

Discussion

Gene regulation patterns in archaea have been intensely studied since the discovery of this new class of organisms in 1977 (Woese and Fox 1977). Because of the difference

among individual archaeal genomes and even among different genes, the gene regulation mechanisms do not seem to fall exclusively into eubacterial or eukaryotic classes. In this paper, we describe an interactive clustering technique that combines two classic methods, PWM and *k*-means clustering algorithms. We applied this technique to four extremophile archaeal genomes. To further examine patterns, we divided ORFs into classes based on location of the ORF with respect to other ORFs. This was an attempt to globally identify genes that were internal members of operons (identified as being ''nearby'' genes) or were the first genes in operons or independent genes (identified as ''distant'' genes). This is clearly only a first approximation; nevertheless, this division did yield useful information. The cutoff chosen for nearby and distance genes was established from histograms of ORF locations with respect to other ORFs. In a similar manner, Slupska et al. (2001) established a cutoff of ± 50 nucleotides in their study of *Pyrobaculum aerophilum*. In order for a cluster to be evident, there must have been a sufficient number of genes that possessed similar sequences in reasonably similar locations within the windows used in the iterative algorithm.

Alignment of archaeal ORFs showed some surprising features that were not related to translation initiation. Each of the four archaea examined showed some evidence of potential transcriptional regulatory sequences in a constant location with respect to the start codon. This is not typical of most eubacteria and eukaryotes.

With the exception of *Halobacterium* sp. NRC-1, most genes examined showed three patterns. Pattern I presented only a G box. These genes have presumptive Shine–Dalgarno sequences in the 5′ flanking region but show no obvious common transcriptional regulatory sequences in a constant location. Pattern II presented only an A box. This is a presumptive TATA sequence in reasonably precise location with location to the start codon. This conserved signal represents a TATA box centered at −26/−27, transcription factor B recognition element (BRE) and 2–3 As in positions at −35 to −33 (Soppa 1999b). These genes show no evidence of a standard ribosome-binding site and so must initiate translation via a leaderless mechanism (Kozak 1999). Pattern III combines an A box and a G box in the common group of genes. These must have both a regulatory site for transcription initiation at a common site and a Shine–Dalgarno site. Table 2 provides a listing of clusters for the various archaea.

Classification as distant or nearby genes also provides information. Genes that are close to one another or even overlap are potential members of the same operon (Salgado et al. 2000). For *Sulfolobus solfataricus* and *Pyrococcus abyssi*, a large number of the pattern I genes are in the nearby class. If they are internal to the operon, they are likely transcribed from some A box upstream to the operon. The G box indicates that they have the potential for internal translation initiation (Salgado et al. 2001; Ma et al. 2002). Genes that are in the distant group but with only a G box most likely have their TATA boxes at some variable distance from the start codon, and so no A box is apparent in the weight matrix.

All of the ORFs in *S. solfataricus* and *P. abyssi* seem to fall into regular categories. On the other hand, *P. aerophilum* ORFs have either an A box in a regular location (nodes 4 and 5) or have no identifiable regular pattern (nodes 6 and 7). While the transcription start might be variable in location, a ribosome-binding site is sufficiently constrained by the nature of translation initiation (Schurr et al. 1993; Ma et al. 2002) that the presence of any should be apparent in the weight matrix. This organism apparently uses leaderless translation initiation exclusively following the eukaryotic pattern (Kozak 1999) and similar to two mycoplasma species, *Mycoplasma genitalium* and *M. pneumoniae* (Boyle and Boyle 2003).

Halobacterium sp. NRC-1 presents an unusual case. The weight matrix for all ORFs shows features not seen in other organisms (Boyle and Boyle 2003). Like *P. aerophilum*, there are no apparent G boxes, suggesting leaderless translation. In addition, there are prominent, conserved peaks of As over a narrow range of sequence locations in nodes 4 and 5. These features are present only in the clusters that have an A box. The A peaks fit no known pattern of 5′ *cis* regulatory sites (Kozak 1999). This could be because the position of this motif relative to the start codon does not exhibit enough regularity to be detected by our method.

The features identified in these four archaeal species reflect their divergence of gene regulation. Using an interactive clustering algorithm, we demonstrated a eukaryotic-type leaderless translation and a bacterial-type leadered translation, which supports the idea that

Table 2 Summary of clustering results. *N* Nearby group, *D* distant group

	(I) G box	(II) A box	(III) A, G box	Other
S. solfataricus (2,977 ORFs)	569[a](N)(6)[b] 737(D)(7)	1383(D)(5) –	288(N)(4) –	–
P. abyssi (1,765 ORFs)	536(N)(6) 265(D)(9)	–	692(N)(7) 427(D)(8)	–
P. aerophilum (2,605 ORFs)	–	317(N)(4) 1332(D)(5)	–	525(N)(6) 431(D)(7)
Halobacterium sp. NRC-1 (2,058 ORFs)	–	420(N)(4) 715(D)(5)	–	84(N)(6) 839(D)(7)

[a]Number of genes in cluster
[b]The node on figures for this organism

archaea have gene regulation machinery with characteristics similar to both prokaryotes and eukaryotes (Saito and Tomita 1999; Soppa 1999a; Tolstrup et al. 2000; Slupska et al. 2001). Some archaeal genes are shown to have transcription initiation elements located at a constant separation from the start of translation. However, several groups of genes with special patterns were also found. The transcription initiation and translation initiation patterns in these groups may reflect special gene regulation patterns in at least some of the archaea species, e.g., *Halobacterium* sp. NRC-1.

The results obtained by our interactive clustering algorithm agree with results previously reported for *S. solfataricus* and *P. abyssi* and reveal potential new regulatory patterns in *Halobacterium* sp. NRC-1. Regulatory signals that are in positions that vary substantially from the modal position may not be recognized by this method because they do not align with similar signals relative to the start codon. Despite this limitation, the clustering algorithm appears to provide valuable qualitative information about sequences such as the Shine–Dalgarno sequence that is expected to have a fairly consistent location relative to the start codon. In addition, it has the power to reveal other unexpected regular features such as the putative TATA box and BRE located at a relatively fixed distance from translation start in a significant subset of ORFs. Although the major focus of the analysis reported in this paper is the identification of the TATA box and Shine–Dalgarno sequences, the real benefit of our technique is the potential to help scientists identify signals in a genome that differ from those reported in other genomes. The confirmation of previously reported patterns and the revelation of patterns that have not been previously described demonstrate the utility of the interactive clustering algorithm for mining new patterns that can be studied by more targeted methods such as those used by Ma et al. (2002) to study Shine–Dalgarno sequences. In the future, we plan to apply the interactive clustering algorithm to other archaea genomic data and further explore the variation of gene regulation of the archaeal kingdom.

Acknowledgements We thank two anonymous reviewers for their critical suggestion for the revision of the initial manuscript.

References

Barns SM, Delwiche CF, Palmer JD, Pace NR (1996) Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci USA 93:9188–9193

Bell SD, Cairns SS, Robson RL, Jackson SP (1999) Transcriptional regulation of an archaeal operon in vivo and in vitro. Mol Cell 4:971–982

Boyle AP, Boyle JA (2003) Global alignment of microbial translation initiation regions. J MS Acad Sci 48:138–150

Bucher P (1990) Weight matrix description of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. J Mol Biol 212:563–578

Dahlke I, Thomm M (2002) A *Pyrococcus* homolog of the leucine-responsive regulatory protein, LrpA, inhibits transcription by abrogating RNA polymerase recruitment. Nucleic Acids Res 30:701–710

DeLong EF, Wu KY, Prezelin BB, Jovine RV (1994) High abundance of Archaea in Antarctic marine picoplankton. Nature 371:695–697

Fitz-Gibbon ST, Ladner H, Kim UJ, Stetter KO, Simon MI, Miller JH (2002) Genome sequence of the hyperthermophilic crenarchaeon *Pyrobaculum aerophilum*. Proc Natl Acad Sci USA 99:984–989

Han J, Kamber K (2000) Data mining: concepts and techniques. Morgan Kaufmann, San Francisco

Holberton DV, Marshall J (1995) Analysis of consensus sequence patterns in *Giardia* cytoskeleton gene promoters. Nucleic Acids Res 15:2945–2953

Jain K, Murthy MN, Flynn PJ (1999) Data clustering: a review. ACM Comput Surv 31:264–323

Kozak M (1999) Initiation of translation in prokaryotes and eukaryotes. Gene 234:187–208

Kyrpides NC, Ouzounis CA (1995) The eubacterial transcriptional activator Lrp is present in the archaeon *Pyrococcus furiosus*. Trends Biochem Sci 20:140–141

Kyrpides NC, Ouzounis CA (1999) Transcription in archaea. Proc Natl Acad Sci USA 96:8545–8550

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

Levy S, Hannenhalli S, Workman C (2001) Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics 17:871–877

Liu R, Blackwell TW, States DJ (2001) Conformational model for binding site recognition by the *E. coli* MetJ transcription factor. Bioinformatics 17:622–633

Ma J, Campbell A, Karlin S (2002) Correlations between Shine–Dalgarno sequences and gene features such as predicted expression levels and operon structures. J Bacteriol 184:5733–5745

Natale DA, Shankavaram UT, Galperin MY, Wolf YI, Aravind L, Koonin EV (2000) Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (COGs). Genome Biol 1:Research0009.1–0009.19

Ng WV, Ciufo SA, Smith TM, Bumgarner RE, Baskin D, Faust J et al (1998) Snapshot of a large dynamic replicon in a halophilic archaeon: megaplasmid or minichromosome? Genome Res 8:1131–1141

Ng WV, Kennedy SP, Mahairas GG, Berquist B, Pan M, Shukla HD et al (2000) Genome sequence of *Halobacterium* species NRC-1. Proc Natl Acad Sci USA 97:12176–12181

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

Saito R, Tomita M (1999) Computer analyses of complete genomes suggest that some archaebacteria employ both eukaryotic and eubacterial mechanisms in translation initiation. Gene 238:79–83

Salgado H, Moreno-Hagelsieb G, Smith TF, Collado-Vides J (2000) Operons in *Escherichia coli*: genomic analyses and predictions. Proc Natl Acad Sci USA 97:6652–6657

Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F et al (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. Nucleic Acids Res 29:72–74

Schurr T, Nadir E, Margalit H (1993) Identification and characterization of *E. coli* ribosomal binding sites by free-energy computation. Nucleic Acids Res 21:4019–4023

She Q, Singh RK, Confalonieri F, Zivanovic Y, Allard G, Awayez MJ et al (2001) The complete genome of the crenarchaeon *Sulfolobus solfataricus* P2. Proc Natl Acad Sci USA 98:7835–7840

Slupska MM, King AG, Fitz-Gibbon S, Besemer J, Borodovsky M, Miller JH (2001) Leaderless transcripts of the crenarchaeal hyperthermophile *Pyrobaculum aerophilum*. J Mol Biol 309: 347–360

Soppa J (1999a) Transcription initiation in archaea: facts, factors and future aspects. Mol Microbiol 31:1295–1305

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

Staden R (1984) Measurements of the effects that coding for a protein has on a DNA sequences and their use for finding genes. Nucleic Acids Res 12:551–567

Tolstrup N, Sensen CW, Garrett RA, Clausen IG (2000) Two different and highly organized mechanisms of translation initiation in the archaeon *Sulfolobus solfataricus*. Extremophiles 4:175–179

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

Wan X, Bridges SM, Boyle JA, Boyle AP (2002) Interactive clustering for exploration of genomic data. In: Dagli CH, Buczak, AL, Ghosh J, Embrechts M, Ersoy O, Kercel S (eds) Smart engineering design, vol 12. ASME Press, New York, pp 753–758

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci USA 74:5088–5090