Received: 15 July 2021 | Revised: 27 July 2021 | Accepted: 29 July 2021

DOI: 10.1002/mbo3.1230

ORIGINAL ARTICLE

MicrobiologyOpen Open Access WILEY

Characterization of promoters in archaeal genomes based on DNA structural parameters

Gustavo Sganzerla Martinez[1] | Sharmilee Sarkar[2] | Aditya Kumar[2] | Ernesto Pérez-Rueda[3] | Scheila de Avila e Silva[1]

[1]Programa de Pós-Graduação em Biotecnologia, Universidade de Caxias do Sul, Caxias do Sul-RS, Brasil

[2]Department of Molecular Biology and Biotechnology, Tezpur University, Tezpur, Assam, India

[3]Unidad Académica de Yucatán, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de México, Mérida, Yucatán, México

Correspondence
Scheila de Avila e Silva, Rua Francisco Getúlio Vargas, 1130, Petrópolis. Caxias do Sul, RS 95070-560, Brazil.
Email: sasilva6@ucs.br

Funding information
Universidade de Caxias do Sul; Junior Research Fellowship; Universidad Nacional Autónoma de México (UNAM), Grant/Award Number: IN-209620; CAPES; Department of Biotechnology, Govt. of India; Department of Biotechnology (DBT), Govt. of India

Abstract

The transcription machinery of archaea can be roughly classified as a simplified version of eukaryotic organisms. The basal transcription factor machinery binds to the TATA box found around 28 nucleotides upstream of the transcription start site; however, some transcription units lack a clear TATA box and still have TBP/TFB binding over them. This apparent absence of conserved sequences could be a consequence of sequence divergence associated with the upstream region, operon, and gene organization. Furthermore, earlier studies have found that a structural analysis gains more information compared with a simple sequence inspection. In this work, we evaluated and coded 3630 archaeal promoter sequences of three organisms, *Haloferax volcanii*, *Thermococcus kodakarensis*, and *Sulfolobus solfataricus* into DNA duplex stability, enthalpy, curvature, and bendability parameters. We also split our dataset into conserved TATA and degenerated TATA promoters to identify differences among these two classes of promoters. The structural analysis reveals variations in archaeal promoter architecture, that is, a distinctive signal is observed in the TFB, TBP, and TFE binding sites independently of these being TATA-conserved or TATA-degenerated. In addition, the promoter encountering meth od was validated with upstream regions of 13 other archaea, suggesting that there might be promoter sequences among them. Therefore, we suggest a novel method for locating promoters within the genome of archaea based on DNA energetic/structural features.

KEYWORDS
archaea, energetic features, structural features, TFBS, transcription

1 | INTRODUCTION

Archaea represent the third domain of life (Woese, 1987) and include an essential and vast variety of organisms with a large diversity of habitats and lifestyles. This cellular domain has many family divisions belonging to four superphyla: TACK, ASGARD, DPANN, and Euryarchaeota. However, well-known information is only available for two divisions, Euryarchaeota and Crenarchaeota, the later being a member of the TACK superphylum. In recent years, with the advent of next-generation sequencing, the availability of archaeal genomes has increased, and more than 300 archaeal genomes have become available to the scientific community, allowing the exploration of

diverse functional and evolutionary mechanisms, such as membrane origin, operon organization, and the proteins devoted to regulating gene expression. Nevertheless, there is a lack of well-annotated archaeal genomic data (Zuo et al., 2015), which enables a lush path toward genomic annotation such as regulatory sequences validation.

The transcription of DNA into RNA and its regulation are central processes in the genetic information flux. Research accumulated in the last few years has evidenced that transcription in archaeal organisms can be roughly described as a simplified version of its eukaryotic relatives (Gehring et al., 2016). The initiation process begins with the binding of a TATA-binding protein (TBP) and a transcription factor B (TFB) to a specific DNA segment, defined as a promoter, allowing the recruitment of the RNA polymerase (RNAP) enzyme. Additionally, the initiation might be optimized with the presence of a transcription factor E (TFE) protein (Ao et al, 2013). Subsequently, an open complex is assembled, followed by the elongation process whereby the RNAP carries out the synthesis of a messenger RNA molecule (mRNA) (Smollet et al., 2017; Soppa, 1999). In general, three main conserved DNA elements devoted to the transcription process have been identified as common to all archaeal groups: (i) an initiator element (INR) around the transcription start site (TSS); (ii) the TATA box element, centered around −26/27 relative to the TSS; and (iii) an element upstream the TATA box comprising two adenines at −34 and −33, which is designated as "transcription factor B recognition element" (BRE). These elements, INR, TATA box, and BRE, are crucial to initiating transcription in archaeal genes. They also present a close homology to eukaryotic transcriptional machinery (Gehring et al., 2016; Soppa, 1999).

An in-depth analysis of archaeal promoter elements will provide comprehension of the gene functionality. As an example, there are advances in biotechnology that have employed promoter identification tools to enhance gene regulation and optimize biological processes. The broader comprehension of promoter activity could, in theory, enable full control over the start and halt of the expression of specific genes (Kernan et al., 2017). The production rise in biosynthetic processes is related to the control of regulatory pathways (Ren et al., 2020). For example, clinical biology has benefited from promoter identification due to the increased mutation rate found in regulatory regions that may lead to antibiotic resistance. Evolutionary biology has also applied promoter identification as part of the process to understand better horizontal gene transfer between species of the three domains of life (Khademi et al., 2019).

Bioinformatics tools employ physical assets of the genetic material and relate these with gene expression variance, enabling the distinction of specific regions such as promoters. The study of DNA structural features may give rise to more information about promoter activity than a primary sequence analysis (Bansal et al., 2014; de Avila e Silva et al., 2011; Kanhere & Bansal, 2005; Yella & Bansal, 2017). Indeed, comparative analysis of bacterial and eukaryotic promoters has shown that Pribnow and TATA boxes, respectively, differ at structure and sequence level from other random locations within and around the promoter (de Avila e Silva et al., 2011; Yella et al., 2018).

When converted into numeric attributes, genetic information will promote enough sensibility for capturing even the smallest alterations among the nucleic acids (Benham, 1996). Hence, we consider the nucleotide conservation found in archaeal promoters (Gribaldo & Brochier-Armanet, 2006; Londei, 2005) will convey a sustained structural parametrization, enabling the characterization of archaeal promoters. In this work, we selected four structural parameters, namely, DNA duplex stability, enthalpy, curvature, and bendability, which are fundamental in understanding the molecular recognition that happens at a structural level (Ryasik et al., 2018).

2 | DATASETS AND METHODS

2.1 | Archaea promoter sequences

To determine the nucleotide composition, a total of 3630 promoter sequences of three archaeal organisms were evaluated, which are divided into 1340 sequences of *Haloferax volcanii* (Babski et al., 2016), 1248 of *Thermococcus kodakarensis* (Jäger et al., 2014), and 1042 of *Sulfolobus solfataricus* (Wurtzel et al., 2009). These particular archaea were selected because they are model organisms and well-studied members of *Halobacteriales*, *Thermococcales*, and *Sulfolobales*, respectively. They also have available transcriptome data (RNAseq), enabling the possibility of retrieving promoter sequences from their published information. Internal and antisense promoters from the transcriptome dataset were not included due to the limitation of data.

The original data covers 1000 nucleotide length sequences, which contains experimentally identified promoters with their transcription start site (TSS), spanning from −500 to +500. Only primary TSS (pTSS) was considered, a category that accounts for abundant transcripts from this original dataset. A shorter sequence was selected, located at 80 nucleotides upstream and 20 nucleotides downstream of the TSS, that is, the core promoter. This briefer region was chosen because it contains the core promoter element (Aptekman & Nadra, 2018; Haberle & Stark, 2018; Kadonaga, 2012). Accordingly, the core promoter has been detailed as sufficient to convey archaeal and eukaryotic transcription (Bartlett et al., 2000; Haberle & Stark, 2018; Zuo et al., 2015). Indeed, promoters from halophilic archaea were reported to be located in the range proposed here; their TATA boxes were found in a median distance of 31 base pairs (bps) upstream the TSS (Babski et al., 2016). Additionally, 96% of the pTSS TATA boxes from *T. kodakarensis* are located in a median distance of 30 base pairs upstream of the TSS (Jäger et al., 2014). The TATA boxes identified in *S. solfataricus* were found in a median length of 35 base pairs upstream of the TSS (Le et al., 2017). Each archaeal promoter sequence had a shuffled version assigned to have a control sequence. The shuffling process was performed by the Supplementary Script S4 (https://doi.org/10.5281/zenodo.5137597).

Moreover, upstream regions from 13 other archaea found in the RSAT Prokaryote Database (Nguyen et al., 2018) were selected to validate the method formulated upon the experimentally verified promoters. *Aciduliprofundum boonei* (741 sequences), *Archaeoglobus*

fulgidus (866 sequences), *Ferroplasma acidarmanus* (430 sequences), *Haloarcula marismortui* (1998 sequences), *Methanocaldococcus jannaschii* (1866 sequences), *Methanosarcina mazei* (822 sequences), *Methanospirillum hungatei* (1467 sequences), *Methanothermobacter thermautotrophicus* (1870 sequences), and *Pyrococcus furiosus* (1286 sequences) were selected as members of Euryarchaea. The following members of TACK archaea were selected: *Caldivirga maquilingensis* (1669 sequences), *Hyphertermus butylicus* (764 sequences), *Ignicoccus*. *hospitalis* (1005 sequences), and *Thermofilum pendens* (1926 sequences). DPANN and ASGARD archaea were not included due to their data unavailability. These particular organisms were selected because of their key role in the evolution of archaea, posing as unique organisms in the archaeal tree of life (Williams et al., 2017).

2.2 | Conversion in structural parameters

To convert the DNA sequences into structural parameters, four DNA structural features were selected, namely DNA duplex stability (DDS), enthalpy contribution, bendability, and intrinsic curvature. These features are biologically relevant to characterize promoter regions since they convert DNA information into numeric attributes (Benham, 1996). These four parameters have previously been used and reflect in capturing specific signals that are not evident at the sequence level (Bansal et al., 2014; de Avila e Silva et al., 2011; Kanhere & Bansal, 2005; SantaLucia & Hicks, 2004; Yella & Bansal, 2017; Yella et al., 2018). Moreover, the appointed features can be described as:

- The DDS of double-stranded DNA is calculated as the sum of its base-pair free energy. It considers the free-energy values associated with the 16 possible combinations of dinucleotides (Kanhere & Bansal, 2005).
- Enthalpy parameters refer to thermodynamic processes that occur at a cellular level (e.g., chemical bonds, mass transport inside and outside the cell, and heat spawning) that affect the thermostability of the cell (Privalov & Crane-Robinson, 2018). These numeric parameters have been taken from DNA melting studies (SantaLucia & Hicks, 2004).
- DNA bendability is a sequence-dependent measurement, reflecting in the DNA bending itself because of the effect specific proteins have in the molecule's helical structure. By this means, DNA bending facilitates the assembly of transcription complexes (Leonard et al., 1997). TATA's bend angle is wider than GC-rich sequences; for instance, TA dinucleotides angle the DNA at 6.74°, the most impactful of the 16 dinucleotide combinations (Karas et al., 1996).
- Finally, intrinsic curvature reflects the capacity of DNA to form small circles around its helical axis (Bolshoy et al., 1991). To this end, we used a model based on DNA gel retardation values (BMHT) for its sensibility toward AT-rich sequences (Bolshoy et al., 1991; Kanhere & Bansal, 2003). BMHT calculation estimated 16 roll and tilt wedge angles based on independent gel mobility experiments performed on a training set of 54 different sequences (Bolshoy et al., 1991).

All the four features selected are sequence-dependent and their combination yields more information gathered on a sequence (Ryasik et al., 2018). The complete set of promoter sequences was converted into structural parameters through a self-developed Python script (Supplementary Script S1: https://doi.org/10.5281/zenodo.5137597) that adopts the numeric parameters available in Table 1, except for intrinsic curvature. The curvature calculation hinges on five nucleotides (instead of di and resulted in 4^5 (Smollet et al., 2017) possible combinations). The 1024 numeric parameters are the result of BMHT calculations (Bolshoy et al., 1991), and they are available in Supplementary Script S2 (https://doi.org/10.5281/zenodo.5137597).

The structural properties were computed in a one-nucleotide sliding window. All promoters were aligned relative to their TSS, and numerical values were averaged to get information in each position.

2.3 | Classification of conserved TATA and degenerated TATA sequences

To classify the core promoters in conserved and degenerated TATA, the MEME Suite—a motif-based sequence analysis tool (Bailey et al., 2009) was employed. All the sequences were scanned with MEME, and the motifs identified by it were extracted. A key motif for this research would be located in −27/-28, so the search was directed to this specific region to capture the TATAs. The following parameters on MEME were used in the organisms *H*. *volcanii* and *T*. *kodakarensis*: *i*) 100 nucleotides sequence length, considering the −80 to +20 region, where the core promoter is located (Haberle & Stark, 2018; Kadonaga, 2012); *ii*) a 0-order background model generated from

TABLE 1 Enthalpy, stability, and bendability parameters for every possible dinucleotide combination

Dinucleotide	Enthalpy (kcal/mol-bp−1)	Stability (kcal/mol-bp−1)	DNA bendability (degrees)
AA	−7.6	−1.00	3.07
AT	−7.2	−0.88	2.6
AC	−8,5	−1.45	2.97
AG	−8.2	−1.3	2.31
TT	−7.6	−1	3.07
TA	−7.2	−0.58	6.74
TC	−7.8	−1.28	2.51
TG	−8.4	−1.44	3.58
CC	−8	−1.28	2.16
CA	−8.5	−1.45	3.58
CT	−7.8	−1.28	2.31
CG	−10.6	−2.24	2.81
GG	−8	−1.84	2.16
GA	−8.2	−1.3	2.51
GT	−8.4	−1.44	2.97
GC	−10.6	−2.24	3.06

the supplied sequences; *iii*) zero or one occurrence (of a contributing motif site) per sequence; *iv*) 8 motifs were located; *v*) the width of the motifs varied between six and eight nucleotides (Hausner et al., 1991). The motif discovery had to follow different parameters in *S. solfataricus*, in which the width of the motifs was increased from six to fifteen nucleotides to capture the TATA boxes adequately. Hence, TATA boxes and BRE elements were considered. The combination of these two consensuses was described as a critical feature in *Sulfolobaceae* family transcription (Le et al., 2017).

Afterward, the dataset was classified through a self-developed Python script (Supplementary Script S3: https://doi.org/10.5281/zenodo.5137597), dividing it into two groups: conserved TATA, those motifs identified by MEME, and degenerated TATA, containing sequences which the previously identified motif was not present.

2.4 | Statistical tests

Statistical tests were conducted to differentiate the two groups this study hinged on. First, the dataset was found not to be normally distributed through the rejection of the null hypothesis

TABLE 2 Conserved TATA and degenerated TATA upon core promoter sequences in three archaeal organisms

		Conserved TATA		Degenerated TATA	
Organism	Genome GC%	Number of promoters (%)	GC%	Number of promoters (%)	GC%
H. volcanii	66.13	21 (1.56%)	54.09	1319 (98.44%)	60.03
T. kodakarensis	50.67	506 (42.72%)	42.55	742 (57.28%)	43.39
S. solfataricus	34.48	840 (80.6%)	28.42	202 (19.4%)	30.68

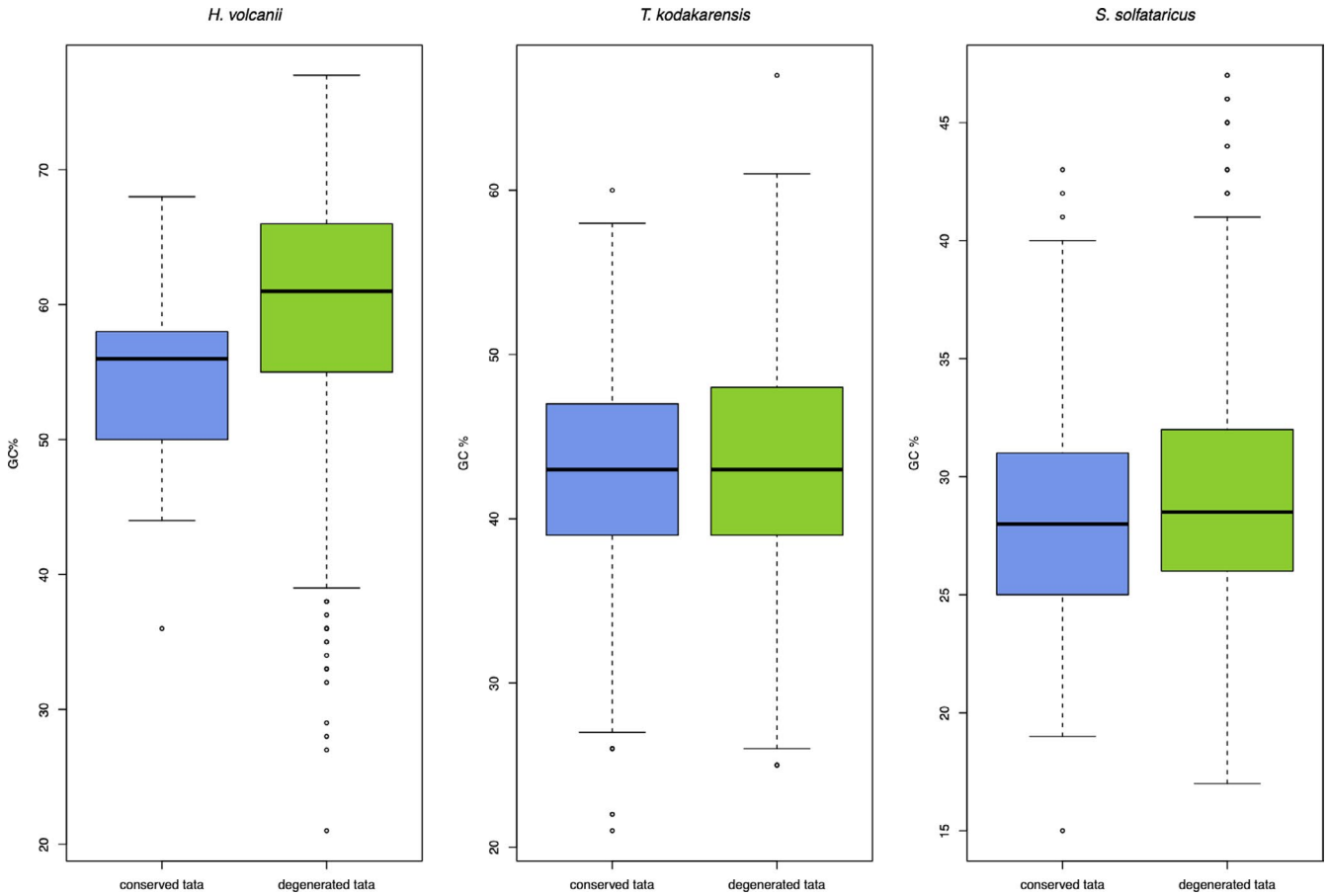

FIGURE 1 Boxplots of TATA-containing and TATA-less promoter sequences in three archaea. We divided 1340 *H. volcanii*, 1248 *T. kodakarensis*, and 1042 *S. solfataricus* sequences into two groups: TATA-containing and TATA-less by following Materials and Methods 2.3. Then, we calculated the GC% of each sequence in the groups and created boxplots alongside U tests to discover significance between the groups. The p values in the nonparametric U tests were as follows: 0.0006556, 0.131, and 3.241e-09 in *H. volcanii*, *T. kodakarensis*, and *S. solfataricus*, respectively

achieved by the Shapiro–Wilk test. Then, to determine if the difference between the groups is significant, the Wilcoxon test was applied. Additionally, the nonparametric Kruskal–Wallis test was conducted to determine the difference between variances in specific organisms. These tests were done in the R programming language in the *stats* package.

3 | RESULTS

3.1 | Sequence composition

The nucleotide composition of the three archaeal organisms was evaluated to denote the genome configuration particular to each archaeon. Firstly, the 1000 nucleotide sequences are composed of 33.8% of AT in *H. volcanii* DS2, 49.3% in *T. kodakarensis* KOD1, and 65.5% in *S. solfataricus* P2. Second, the core promoter elements (−80 to +20) in these organisms presented an AT value of 40% in *H. volcanii*, 56.8% in *T. kodakarensis*, and 71.1% in *S. solfataricus*.

3.2 | Conserved TATA and Degenerated TATA boxes.

The datasets were split into two groups to capture particularities and verify the hypothesis of the archaeal transcription being beyond TATA box conservation. The two groups are Conserved TATA and Degenerated TATA. Motifs of eight nucleotides were found in *H. volcanii* and *T. kodakarensis*. Simultaneously, the outcome of *S. solfataricus* encompassed 14 nucleotides. In an attempt to preserve the particularities each archaeon has, the analysis was individually done. The TATA box motif of each organism is found in Figure A1, from where *H. volcanii* presented SYTTWWAA, *T. kodakarensis* TATA was identified as VYTTWWAA, and *S. solfataricus* accounted for KVRWAAA VYTTWWWW motifs.

When each one of the motifs was employed to split the dataset, the results of Table 2 were produced. To begin with, 1.56% of 1340 *H. volcanii* sequences presented the TATA motif previously identified. The number of sequences containing motifs in *T. kodakarensis* was 42.72% and 80.6% in *S. solfataricus*. Then, the GC% of each group was evaluated to verify if they yield statistical significance. U tests

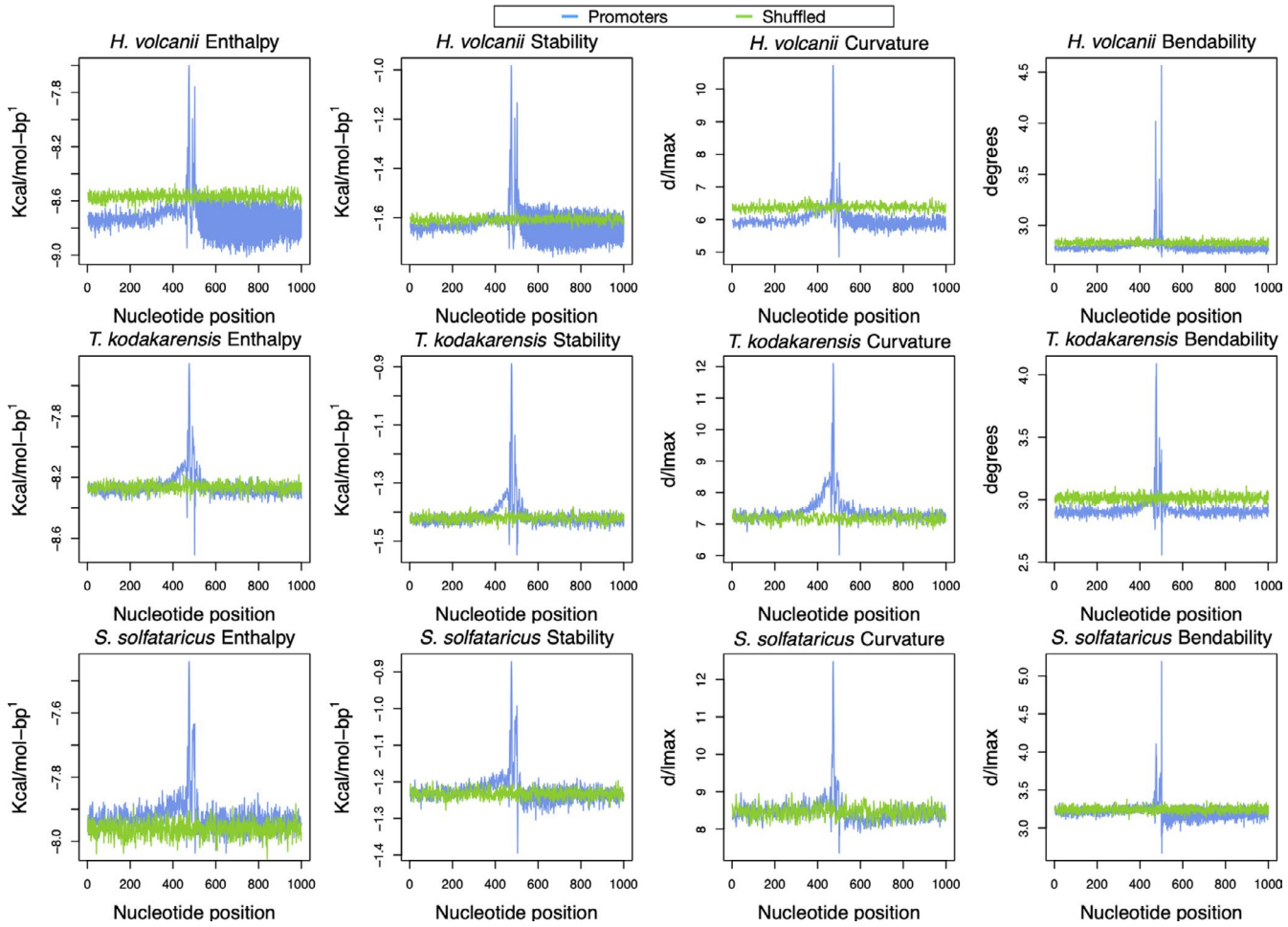

FIGURE 2 Structural/energetic profiles of 1000 nucleotides found in promoter and shuffled sequences. Energetic/structural features of three archaea. We plotted the average value in each one of the 1000 positions. The highest peak is seen at position −28 in three archaea, four measurements. The blue line represents the promoter sequences and the green line indicates a shuffled version of the promoters. The shuffling process was carried out by a Python script (Supplementary Script S4: https://doi.org/10.5281/zenodo.5137597)

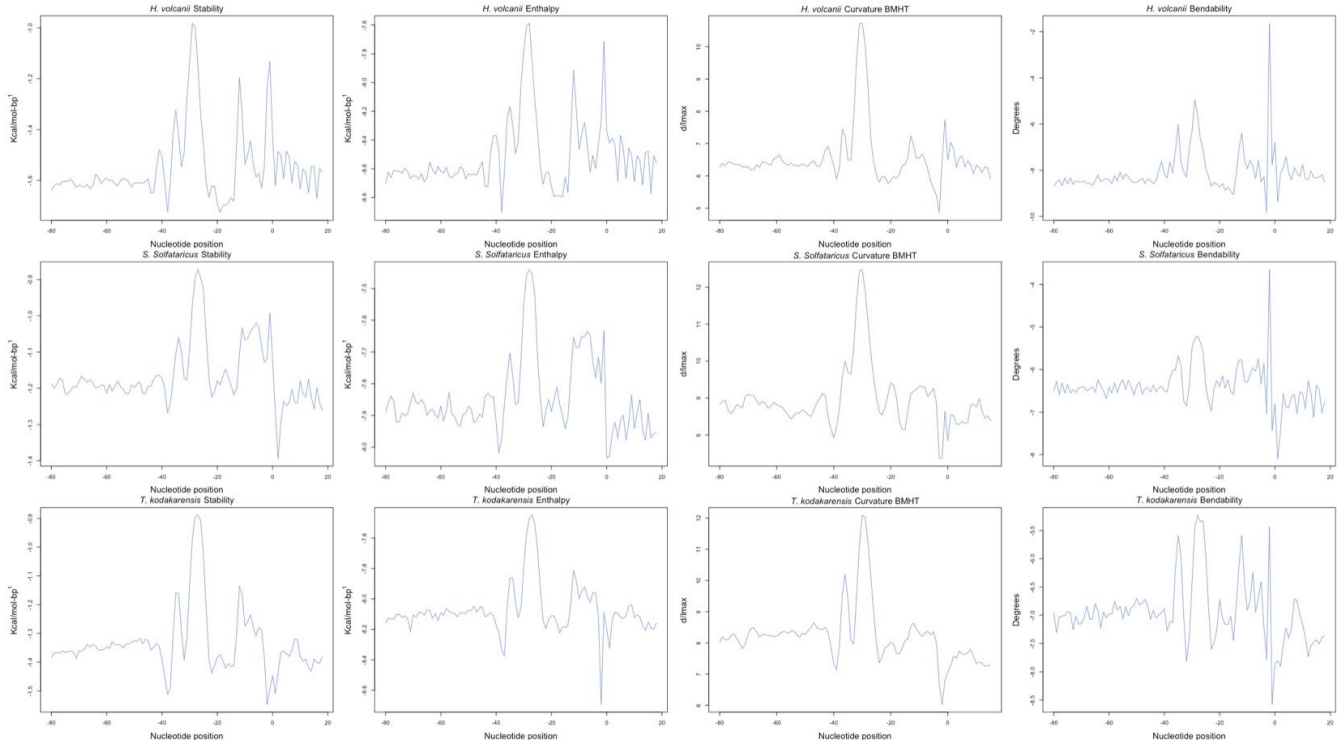

FIGURE 3 Structural/energetic core promoter profiles. Energetic and sequence-dependent features of three archaea. We plotted the average of the core promoter positions reported by Kadonaga, 2012; Haberle and Stark, 2018. Our plots indicated a strong signal in i) the TATA box and BRE positions; ii) the PPE area

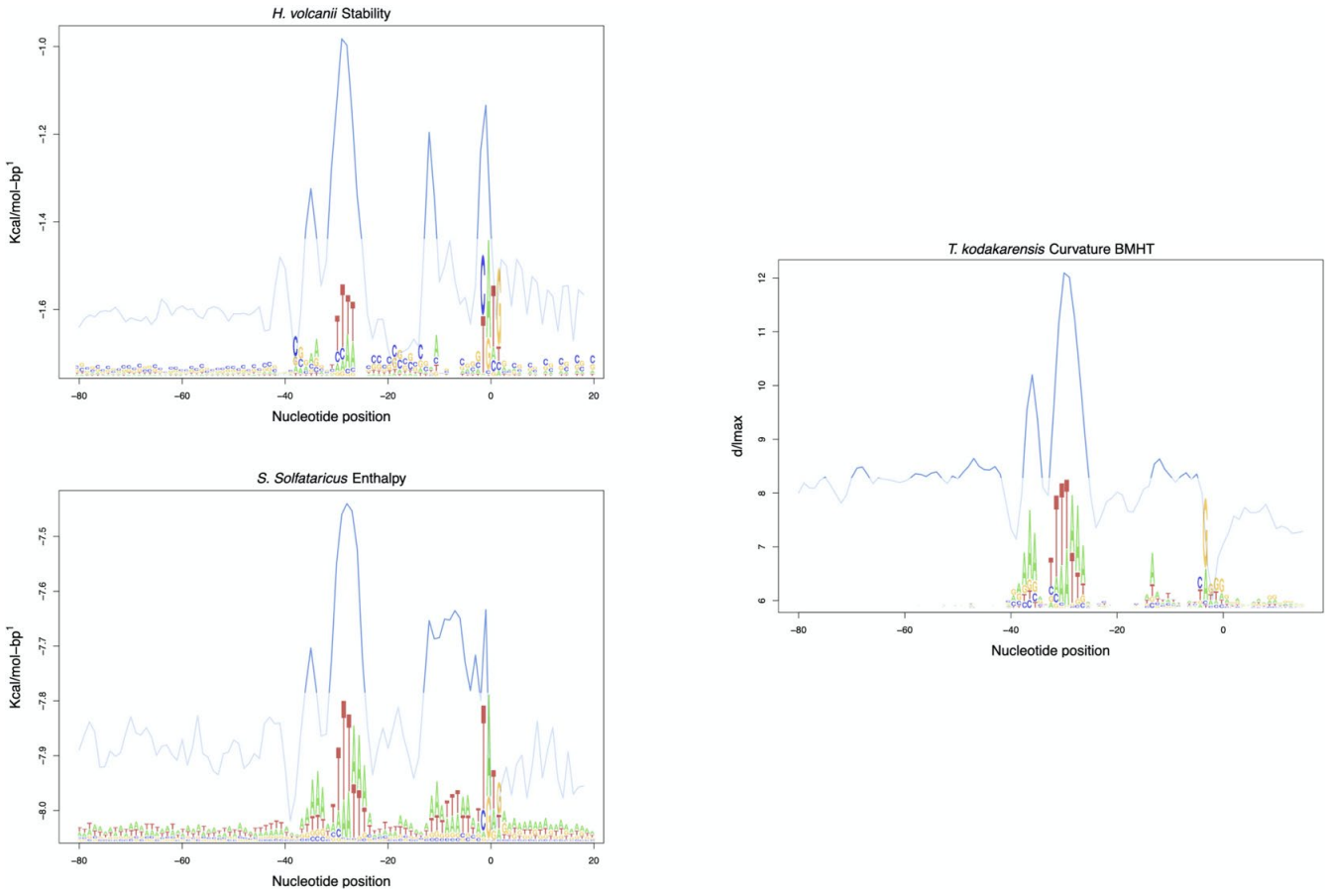

FIGURE 4 Transcription factor binding sites represented by signals regarding structural/energetic profiles of the core promoter. Nucleotide information (sequence logo profiles) is overlaid with signals that represent the core promoter content

were performed due to the data not following a normal distribution. Figure 1 shows boxplots from which the means of the conserved and degenerated TATA in *H. volcanii*, *T. kodakarensis*, and *S. solfataricus* are $p = 0.0006556$, $p = 0.131$, and $p = 0.005365$, respectively.

3.3 | Structural profiles of archaeal promoter sequences vary when transcription factors binding sites

The entire promoter dataset was converted into enthalpy, DNA Duplex Stability (DDS), bendability, and intrinsic curvature to capture specific signals in wider genome analysis, ranging from −500 to +500. In addition, control sequences were added to elicit the strong signals promoter sequences have (Figure 2). A zoomed version, encompassing the promoter region only, was included in Figure 3, where there is a conserved region around the binding site of the transcription factor proteins: TBP (TATA box, around −28), TFB (BRE, around 2 nucleotides upstream TBP), TFE, whose binding site is located in position −10 (PPE – proximal promoter element) and +1, matching the INR (initiator element).

3.4 | Definition of a promoter-like profile

By following the profiles brought by Figure 3, a promoter-like profile was formed upon the average per position (100 nucleotides) of each feature in the validated promoter dataset. By combining nucleotide information (sequence logo profiles) with the structural parametrization brought by this work, Figure 4 was created. In this, the strong DDS, enthalpy, bendability, and BMHT curvature signals are overlaid with transcription factor binding sites.

3.5 | Validation of the results with 13 other archaea

Upstream regions of thirteen other archaea divided into four TACKs and nine *Euryarchaea* were included to test the validity of the findings. Figure 5 holds the genomic information of each archaeon plotted against DNA bendability, BMHT curvature, enthalpy, and DDS. In all cases, a strong signal around the ending of the upstream regions was located.

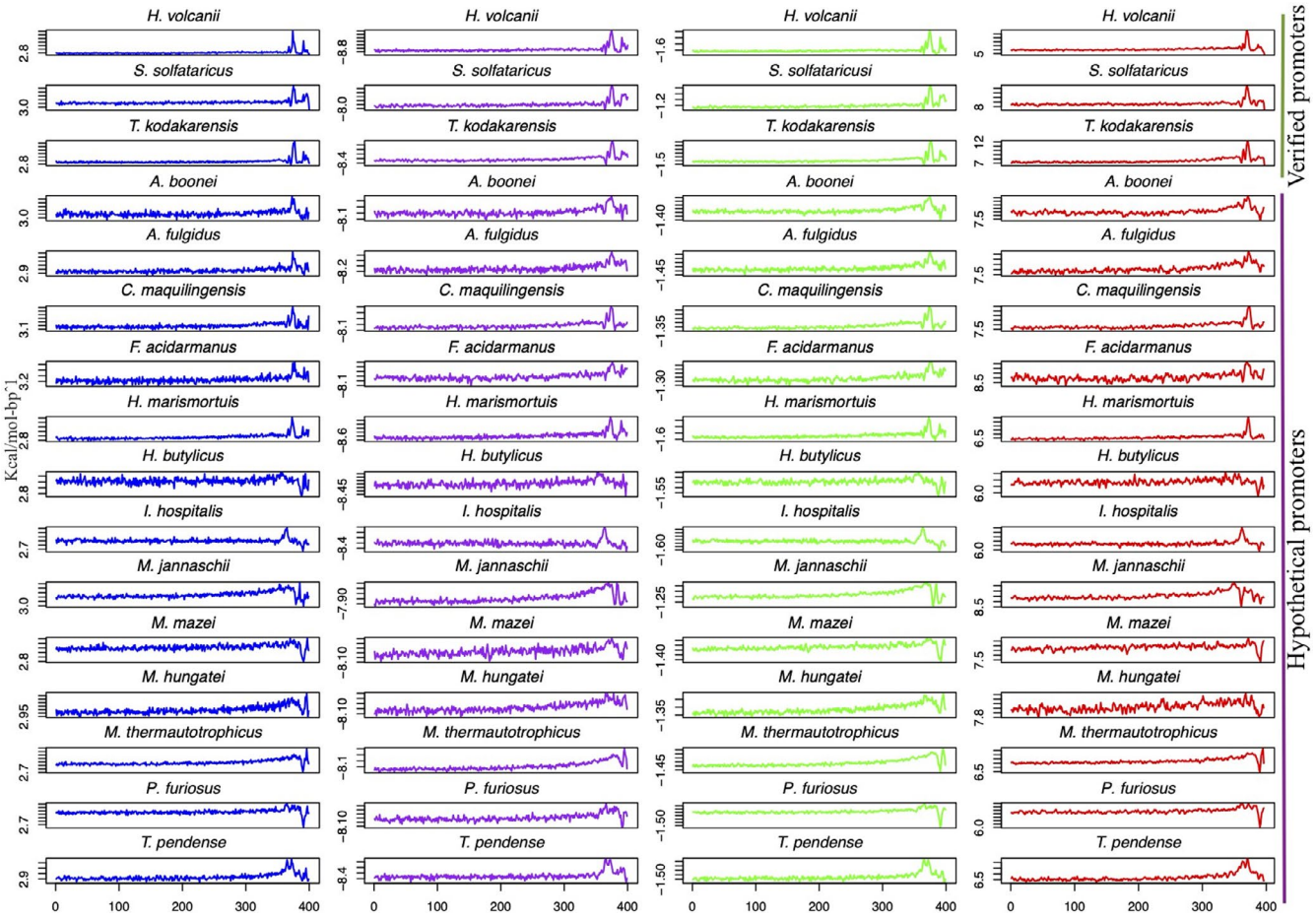

FIGURE 5 Structural/energetic upstream profiles in thirteen archaea. Thirteen other archaea were selected from 42 to validate the promoter-like behavior observed. These organisms have 400 nucleotide-long sequences corresponding to upstream sequences where no annotation toward promoter finding was done. The blue lines represent bendability profiles, the purple enthalpy, the green refers to DDS, and the red is BMHT curvature

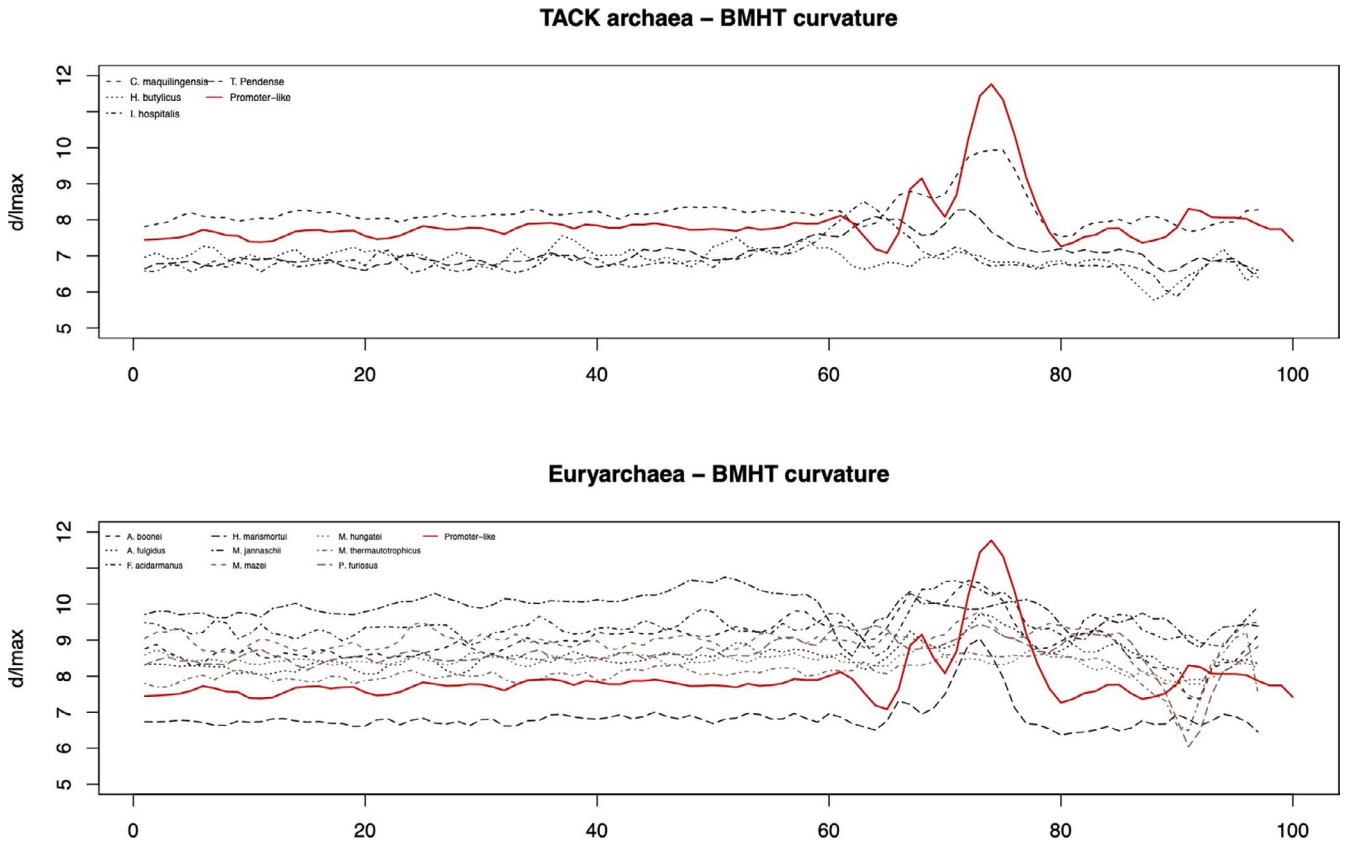

FIGURE 6 Bendability signal comparison of promoters and upstream regions of thirteen other archaea. The red line (promoter-like) represents the average formed upon experimentally validated promoter of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis* to be compared with upstream sequences of thirteen other archaea divided into two phylogenetic families. The remaining DDS, enthalpy, and BHMT curvature are found in Figures A2, A3, and A4, respectively

Moreover, we included a comparison of the upstream regions found in 13 archaea against the promoter-like profile established in 3.4. To perform a comparative analysis, the promoter-like profile was compared with upstream regions of 13 other archaea split into their phylogenetic family (Figure 6). Since the profiles observed in Figure 6 are the same when another physical feature is tested, comparisons following DDS, enthalpy, and bendability are included in Figures A2, A3, and A4, respectively. Analysis of variance tests indicated each organism is significantly different than the other by presenting $p < 2e\text{-}16$ in TACK archaea and $p < 2e\text{-}16$ in *Euryarchaea*. The statistical analysis of the two archaeal families is visualized in boxplots available in Figure 7.

3.6 | Conserved and degenerated TATA groups

The core promoters belonging to conserved and degenerated TATA groups were converted into energetic and structural properties to indicate RNAP action in both groups (Figure 8). The two groups presented overlapping lines with strong signals being located around −28.

4 | DISCUSSION

4.1 | Nucleotide content

The results of this study suggest that TATA boxes slightly vary between organisms, supporting the archaeal diversification reported by (DeLong et al., 1994). Additionally, the AT content was found differently in each archaeon.

When the archaeal promoters were evaluated as owning either a conserved or a degenerated TATA consensus, the GC% of each organism has explained the conservation found upon TATA boxes, so the organism with higher genome GC% was the one that presented the least amount of TATAs, this is no news. However, the binding of TBP, TFB, and TFE to a TATA+BRE motif and TFE binding to PPE/INR were found through this *in silico* approach to be off from a primary sequence inspection, just as that conservation found around these motifs is not mandatory. Moreover, promoter activity is still observed when promoters lack a clear TATA motif (Aptekman & Nadra, 2018). Therefore, the uneven number of conserved TATA sequences sprung around archaea is explained by the dynamics of biology. The two groups of promoters (conserved and degenerated TATA) have

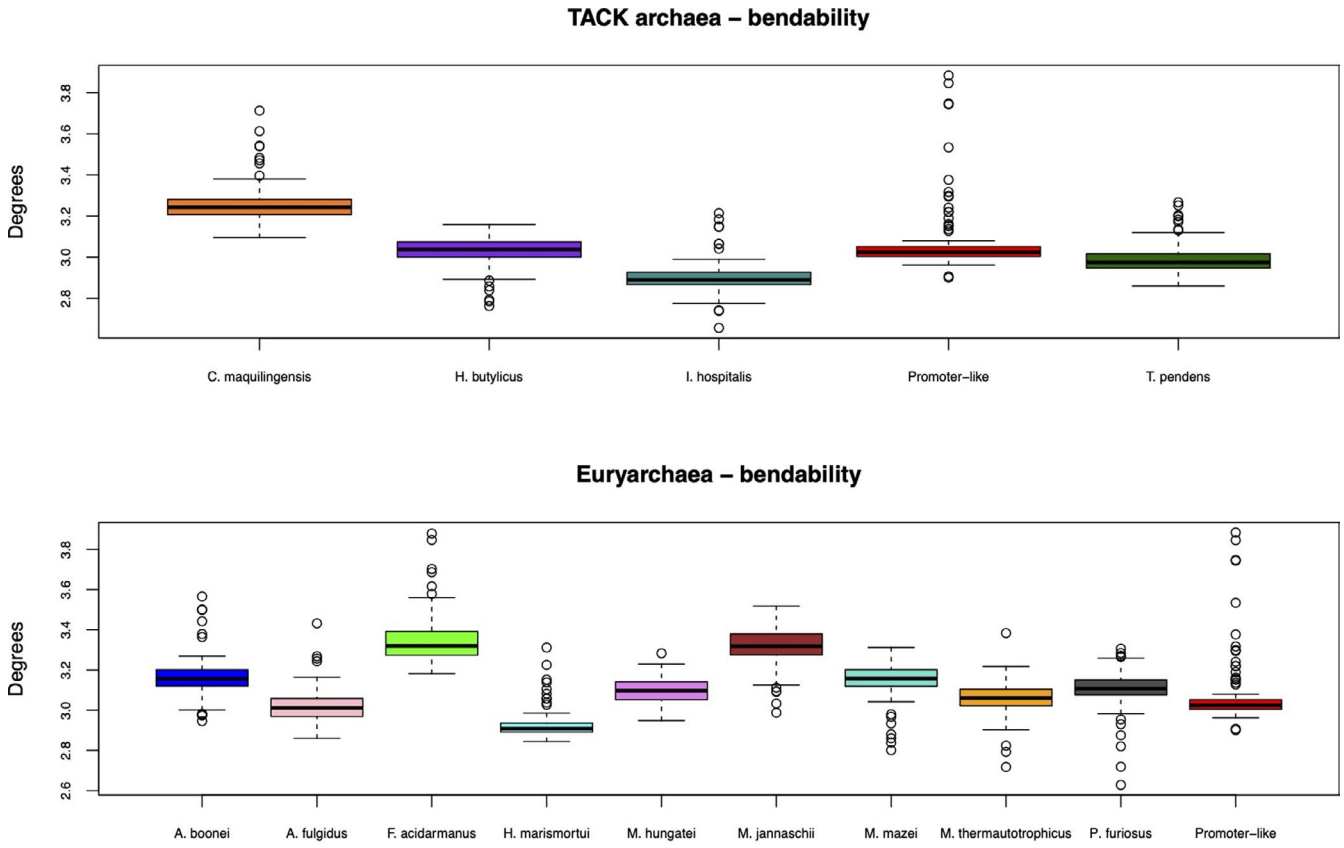

FIGURE 7 Boxplots of promoters and upstream regions of thirteen other archaea converted to bendability. The boxplots represent statistical comparisons between the promoter-like profile, (red), formed upon experimental data of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis*. The $p < 2e\text{-}16$ values obtained by the nonparametric Kruskal–Wallis test conveyed statistical significance in the averages of both groups. Additional analyses encompassing BMHT curvature, enthalpy, and DDS are found in Figures A5, A6, and A7, respectively

also presented statistical significance in *H. volcanii* and *S. solfataricus* when the GC content was employed as a possible explanation for each group. This reassures the hypothesis that the probability of TATA boxes to be found depends directly on the genome composition of a given archaeon.

4.2 | Energetic and structural parameters define promoter-like profiles

Promoter sequences might be defined by a set of strong signals around their transcription factor binding sites (TFBS), that is, TFB, TBP, and TFE. In this study, the conversion of genetic information into physical attributes has protruded distinctive signals around TFBS of the proteins, while shuffled sequences did not. These strong signals are in favor of the relative location of the basal transcription factors (TF), which is explained by the laws ruling the promoter area. Both enthalpy and stability are energetic-related features, the base pairs that are more commonly found in promoters are AT and their chemical conformation reflects in more energy available (Allawi & SantaLucia, 1997; de Avila e Silva et al, 2014; Privalov & Crane-Robinson, 2018; SantaLucia & Hicks, 2004; Yella et al., 2018). The distinct signals represented by curvature and bendability are explained by the TFBS being more rigid and more curved, which acts against the formation of nucleosomes (Tirosh et al., 2007).

The profiles obtained in this study indicate a conserved aspect around the binding site of proteins that are key elements in the Pre-Initiation Complex (PIC) formation. In vitro studies advocated for TBP+TFB being enough to begin transcription. Indeed, our results show conserved signals around this site (−27 2nt spacer −31). However, the inclusion of a signal in the vicinity of −10 and +1, which matches the TFE binding site, also contributes to promoter definition (Ao et al, 2013). This TF protein was reported to optimize the formation of PIC in TACK and other families as well (Hanzelka et al., 2001).

The signal located in the −10 region of three archaea is also an important factor in bacterial transcription (Lloréns-Rico et al., 2015). Both bacteria and archaea share the same last unique common ancestor, and consequently, share similarities despite their evolution taking place in different branches of the tree of life (Gribaldo & Brochier-Armanet, 2006).

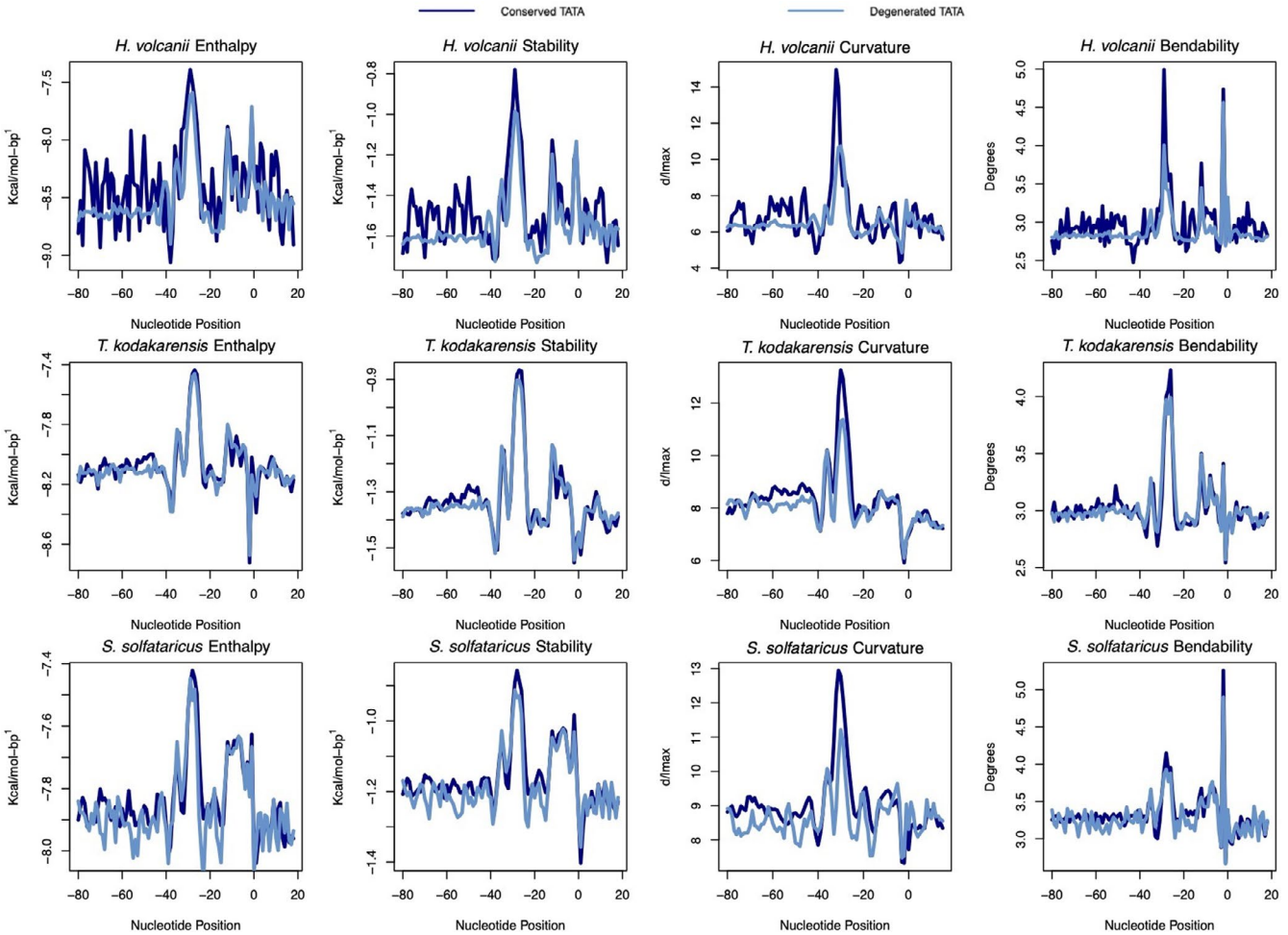

FIGURE 8 Structural/energetic profiles of conserved and degenerated TATA promoters. The conserved and degenerated TATA core promoter profiles are plotted. The lines represent the average value each group and organism showed. The navy-blue lines represent sequences that had a MEME-identified TATA motif, the light blue depicts sequences in which the specific TATA motif was not found

The lack of annotation in the genome of many archaea creates the possibility for such methods. When the validation of the promoter identification method was tested in upstream regions of thirteen archaea, the same rationale was inferred. Mining published information upon transcripts has enabled the definition of a promoter-like profile through a combination of strong signals in the binding sites of TBP, TFB, and TFE (−27, −31, −10, and +1, respectively). When data that do not encompass experimentally validated promoter sequences only was assessed, strong signals were observed in the ending of the sequences, suggesting that there might be promoter elements found in these intergenic areas, as identified by (Yella et al., 2018).

The observation of Figure 6 (and Figures A2, A3, and A4) assures the possibility of locating promoters in upstream regions due to their physical profile. Two archaea have shown TFBS signals similar to the promoter-like profile: *A. boonei* and *T. pendens*. Even though there are differences in the signals protruded by promoters and potential promoters, resulting in significant differences between the groups' averages, the second group poses for the rise of methods for promoter identification as the one brought by this study.

4.3 | Promoter signal beyond TATA boxes

TATA boxes are likely the most conserved sites that distinguish both archaeal/eukaryotic promoters. The initiation of the transcription in archaea has been reported to start with TPB and TFB proteins attaching to the promoter (Gehring et al., 2016, Blombach & Grohmann, 2017) and enhanced by the presence of TFE (Hanzelka et al., 2001), this binding is assisted by the conservation found around the binding site of these proteins. Promoters have been grouped in terms of their TATA analysis in (Tirosh et al., 2007; Yella & Bansal, 2017), both authors performed structural conversions such as this study did. Divergent results could be observed in which TATA-conserved sequences did not show significant differences when compared to TATA-degenerated ones.

In this study, both TATA-conserved and TATA-degenerated groups have shown the same strong signals around the binding sites of TFB, TBP, and TFE. Some differences might protrude mathematical variance, for example, the TFB and TBP binding sites analyzed in the curvature profile of three archaea and *H. volcanii* bendability and DDS. This feature defines the promoter (either TATA-conserved or not) as a promoter-like sequence, which is a novel approach in identifying and finding new promoter sequences in archaea.

5 | CONCLUSIONS

The results we demonstrated in this study encourage the DNA codification into energetic/structural attributes that reveal transcription factor proteins binding sites where a primary sequence inspection failed. Hence, this study poses a novel method to be used in genome annotation regarding archaeal promoters.

ACKNOWLEDGMENTS

We are grateful to the support received from Universidade de Caxias do Sul (UCS), Coordernação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), the Department of Biotechnology (DBT), Govt. of India for project Junior Research Fellowship, and the Department of Biotechnology, Govt. of India for the DBT Twinning project grant (BT/PR24927/NER/95/911/2017). This research was supported by grants from Dirección General de Asuntos del Personal Académico-Universidad Nacional Autónoma de México (UNAM) (IN-209620).

CONFLICT OF INTERESTS

None declared.

AUTHOR CONTRIBUTIONS

Gustavo Sganzerla Martinez: Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing-original draft (equal); Writing-review & editing (equal). **Sharmilee Sarkar:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing-original draft (equal); Writing-review & editing (equal). **Aditya Kumar:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing-original draft (equal); Writing-review & editing (equal). **Scheila de Avila e Silva:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing-original draft (equal); Writing-review & editing (equal). **Ernesto Perez-Rueda:** Conceptualization (equal); Data curation (equal); Formal analysis (equal); Funding acquisition (equal); Investigation (equal); Methodology (equal); Project administration (equal); Resources (equal); Software (equal); Supervision (equal); Validation (equal); Visualization (equal); Writing-original draft (equal); Writing-review & editing (equal).

ETHICS STATEMENT

None required.

DATA AVAILABILITY STATEMENT

The experimentally verified promoter sequences were retrieved from their original publications: H. volcanii (https://doi.org/10.1186/s12864-016-2920-y), S. solfataricus (https://doi.org/10.1101/gr.100396.109), and T. kodakarensis (https://doi.org/10.1016/0022-2836(91)90492-O). The upstream regions used in the method validation step were extracted from RSAT Database (http://www.rsat.eu). The supplementary material is available in the Zenodo repository: (1) the sequence IDs and the gene annotation of all H. volcanii, T. kodakarensis, and S. solfataricus promoters used in this study: https://doi.org/10.5281/zenodo.5137550, (2) the Python scripts S1–S4 employed in the structural analysis of archaeal promoter sequences: https://doi.org/10.5281/zenodo.5137597

ORCID

Gustavo Sganzerla Martinez https://orcid.org/0000-0002-7656-0579
Sharmilee Sarkar https://orcid.org/0000-0001-5655-6874
Aditya Kumar https://orcid.org/0000-0002-6474-8830
Ernesto Pérez-Rueda https://orcid.org/0000-0002-6879-0673
Scheila de Avila e Silva https://orcid.org/0000-0002-3472-3907

REFERENCES

Allawi, H. T., & SantaLucia, J. (1997). Thermodynamics and NMR of internal G-T Mismatches in DNA. *Biochemistry*, *97*(36), 10581–10594. https://doi.org/10.1021/bi962590c

Ao, X., Li, Y., Wang, F., Feng, M., Lin, Y., Zhao, S., Liang, Y., & Peng, N. (2013). The sulfolobus initiation element is an important contributor to promoter strength. *Journal of Bacteriology*, *195*(22), 5216–5222.

Aptekman, A. A., & Nadra, A. D. (2018). Core promoter information content correlates with optimal growth temperature. *Scientific Reports*, *8*, 1313. https://doi.org/10.1038/s41598-018-19495-8

Babski, J., Haas, K. A., Näther-Schindler, D., Pfeiffer, F., Förstner, K. U., Hammelmann, M., Hilker, R., Becker, A., Sharma, C. M., Marchfelder, A., & Soppa, J. (2016). Genome-wide identification of transcriptional start sites in the haloarchaeon Haloferax volcanii based on differential RNA-Seq (dRNA-Seq). *BMC Genomics*, *17*, 629. https://doi.org/10.1186/s12864-016-2920-y

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Jingyuan, R., Wilfred, W. L., & Noble, W. S. (2009). MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Research*, *37*(Web Server), W202–W208. https://doi.org/10.1093/nar/gkp335

Bansal, M., Kumar, A., & Yella, V. R. (2014). Role of DNA sequence based structural features of promoters in transcription initiation and gene

expression. *Current Opinion in Structural Biology*, *25*, 77–85. https://doi.org/10.1016/j.sbi.2014.01.007

Bartlett, M. S., Thomm, M., & Geiduschek, E. P. (2000). The orientation of DNA in an archaeal transcription initiation complex. *Natural Structural Biology*, *7*, 782–785.

Benham, C. J. (1996). Duplex destabilization in superhelical DNA is predicted to occur at specific transcriptional regulatory regions. *Journal of Molecular Biology*, *225*(3), 425–434. https://doi.org/10.1006/jmbi.1996.0035

Blombach, F., & Grohmann, D. (2017). Same same but different: The evolution of TBP in archaea and their eukaryotic offspring. *Transcription*, *8*(3), 162–168. https://doi.org/10.1080/21541264.2017.1289879

Bolshoy, A., McNamara, P., Harrington, R. E., & Trifonov, E. N. (1991). Curved DNA without A-A: Experimental estimation of all 16 DNA wedge angles. *Proceedings of the National Academy of Sciences USA*, *88*(6), 2312–2316.

de Avila e Silva, S., Echeverrigaray, S., & Gerhardt, G. J. L. (2011). BacPP: Bacterial promoter prediction-A tool for accurate sigma-factor specific assignment in enterobacteria. *Journal of Theoretical Biology*, *287*, 92–99. https://doi.org/10.1016/j.jtbi.2011.07.017

de Avila e Silva, S., Forte, F., Sartor, I., Andrighetti, T., Gerhardt, G., Longaray Delamare, A. P., & Echeverrigaray, S. (2014). DNA duplex stability as discriminative characteristic for Escherichia coli σ54- and σ28- dependent promoter sequences. *Biologicals*, *42*(1), 22–28. https://doi.org/10.1016/j.biologicals.2013.10.001

DeLong, E. F., Wu, K. Y., Prézelin, B. B., & Jovine, R. V. M. (1994). High abundance of Archaea in Antarctic marine picoplankton. *Nature*, *371*, 695–697. https://doi.org/10.1038/371695a0

Gehring, A. M., Walker, J. E., & Santangelo, T. J. (2016). Transcription regulation in archaea. *Journal of Bacteriology*, *198*(14), 1906–1917. https://doi.org/10.1128/JB.00255-16

Gribaldo, S., & Brochier-Armanet, C. (2006). The origin and evolution of Archaea: A state of the art. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *361*, 1470. https://doi.org/10.1098/rstb.2006.1841

Haberle, V., & Stark, A. (2018). Eukaryotic core promoters and the functional basis of transcription initiation. *Nature Reviews Molecular Cell Biology*, *19*, 621–637. https://doi.org/10.1038/s41580-018-0028-8

Hanzelka, B. L., Darcy, T. J., & Reeve, J. N. (2001). TFE, an archaeal transcription factor in *Methanobacterium thermoautotrophicum* related to eucaryal transcription factor TFIIEα. *Journal of Bacteriology*, https://doi.org/10.1128/JB.183.5.1813-1818.2001

Hausner, W., Frey, G., & Thomm, M. (1991). Control regions of an archaeal gene. A TATA box and an initiator element promote cell-free transcription of the tRNAVal gene of Methanococcus vannielii. *Journal of Molecular Biology*, https://doi.org/10.1016/0022-2836(91)90492-O

Jäger, D., Förstner, K. U., Sharma, C. M., Santangelo, T. J., & Reeve, J. N. (2014). Primary transcriptome map of the hyperthermophilic archaeon Thermococcus kodakarensis. *BMC Genomics.*, *222*(5), 495–508. https://doi.org/10.1016/0022-2836(91)90492-O

Kadonaga, J. T. (2012). Perspectives on the RNA polymerase II core promoter. *Wiley Interdisciplinary Reviews: Developmental Biology*, *1*(1), 40–51. https://doi.org/10.1002/wdev.21

Kanhere, A., & Bansal, M. (2003). An assessment of three dinucleotide parameters to predict DNA curvature by quantitative comparison with experimental data. *Nucleic Acids Research*, *2647*, 2658.

Kanhere, A., & Bansal, M. (2005). A novel method for prokaryotic promoter prediction based on DNA stability. *BMC Bioinformatics*, *6*, 1. https://doi.org/10.1186/1471-2105-6-1

Karas, H., Knuppel, R., Schuiz, W., Sklenar, H., & Wingender, E. (1996). Combining structural analysis of dna with search routines for the detection of transcription regulatory elements. *Bioinformatics*, *12*(5), 441–446. https://doi.org/10.1093/bioinformatics/12.5.441

Kernan, T., West, A. C., & Banta, S. (2017). Characterization of endogenous promoters for control of recombinant gene expression in *Acidithiobacillus ferrooxidans*. *Biotechnology and Applied Biochemistry*, *64*, 6. https://doi.org/10.1002/bab.1546

Khademi, S. M., Sazinas, P., & Jelsbak, L. (2019). Within-Host adaptation mediated by intergenic evolution in Pseudomonas aeruginosa. *Genome Biology and Evolution*, *11*(5), 1385–1397. https://doi.org/10.1093/gbe/evz083

Le, T. N., Wagner, A., & Albers, S. (2017). A conserved hexanucleotide motif is important in UV-inducible promoters in *Sulfolobus acidocaldarius*. *Microbiology (N Y)*, *163*(5), 778–788. https://doi.org/10.1099/mic.0.000455

Leonard, D. A., Rajaram, N., & Kerppola, T. K. (1997). Structural basis of DNA bending and oriented heterodimer binding by the basic leucine zipper domains of Fos and Jun. *Proceedings of the National Academy of Sciences USA*, *94*(10), 4913–4918. https://doi.org/10.1073/pnas.94.10.4913

Lloréns-Rico, V., Lluch-Senar, M., & Serrano, L. (2015). Distinguishing between productive and abortive promoters using a random forest classifier in Mycoplasma pneumoniae. *Nucleic Acids Research*, *43*(7), 3442–3453. https://doi.org/10.1093/nar/gkv170

Londei, P. (2005). Evolution of translational initiation: New insights from the archaea. *FEMS Microbiology Reviews*, *29*(2), 185–200. https://doi.org/10.1016/j.fmrre.2004.10.002

Nguyen, N. T. T., Contreras-Moreira, B., Castro-Mondragon, J. A., Santana-Garcia, W., Ossio, R., Robles-Espinoza, C. D., & Thomas-Chollier, M. (2018). RSAT 2018: Regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Research*, https://doi.org/10.1093/nar/gky317

Privalov, P. L., & Crane-Robinson, C. (2018). Forces maintaining the DNA double helix and its complexes with transcription factors. *Progress in Biophysics and Molecular Biology*, *135*, 30–48. https://doi.org/10.1016/j.pbiomolbio.2018.01.007

Ren, H., Shi, C., & Zhao, H. (2020). Computational tools for discovering and engineering natural product biosynthetic pathways. *Iscience*, *23*, 1. https://doi.org/10.1016/j.isci.2019.100795

Ryasik, A., Orlov, M., Zykova, E., Ermak, T., & Sorokin, A. (2018). Bacterial promoter prediction: Selection of dynamic and static physical properties of DNA for reliable sequence classification. *Journal of Bioinformatics and Computational Biology*, *16*, 1. https://doi.org/10.1142/S0219720018400036

SantaLucia, J., & Hicks, D. (2004). The Thermodynamics of DNA Structural Motifs. *Annual Review of Biophysics and Biomolecular Structure*, *33*, 415–440. https://doi.org/10.1146/annurev.biophys.32.110601.141800

Smollet, K., Blombach, F., Fouqueau, T., & Wernerm, F. (2017) A Global Characterisation of the Archaeal Transcription Machinery. In: B. Clouet (eds) *RNA metabolism and Gene Expression in Archaea*. Nucleic Acids Mol Biol, 32.

Soppa, J. (1999). Transcription initiation in Archaea: Facts, factors and future aspects. *Molecular Microbiology*, *31*, 5. https://doi.org/10.1046/j.1365-2958.1999.01273.x

Tirosh, I., Berman, J., & Barkai, N. (2007). The pattern and evolution of yeast promoter bendability. *Trends in Genetics*, *23*(7), 318–321. https://doi.org/10.1016/j.tig.2007.03.015

Williams, T. A., Szöllosi, G. J., Spang, A., Foster, P. G., Heaps, S. E., Boussau, B., & Martin Embley, T. (2017). Integrative modeling of gene and genome evolution roots the archaeal tree of life. *Proceedings of the National Academy of Sciences*, *114*(23), E4602–E4611

Woese, C. R. (1987). Bacterial evolution. *Microbiology and Molecular Biology Reviews*, *51*(2), 221–271. https://doi.org/10.1128/MMBR.51.2.221-271.1987

Wurtzel, O., Sapra, R., Chen, F., Zhu, Y., Simmons, B. A., & Sorek, R. (2009). A single-base resolution map of an archaeal transcriptome. *Genome Research*, *20*, 133–141.

Yella, V. R., & Bansal, M. (2017). DNA structural features of eukaryotic TATA-containing and TATA-less promoters. *FEBS Open Biology*, *7*(3), 324–334. https://doi.org/10.1002/2211-5463.12166

Yella, V. R., Kumar, A., & Bansal, M. (2018). Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy.

Scientific Reports, *8*, 4250. https://doi.org/10.1038/s41598-018-22129-8

Yella, V. R., Kumar, A., & Bansal, M. (2018). Identification of putative promoters in 48 eukaryotic genomes on the basis of DNA free energy. *Scientific Reports*, https://doi.org/10.1038/s41598-018-22129-8

Zuo, G., Xu, Z., & Hao, B. (2015). Phylogeny and taxonomy of archaea: A comparison of the Whole-Genome-Based CVtree approach with 16s rRNA sequence analysis. *Life*, *5*(1), 949–968. https://doi.org/10.3390/life5010949

How to cite this article: Martinez, G. S., de Avila e Silva, S., Sarkar, S., Kumar, A., & Pérez-Rueda, E. (2021). Characterization of promoters in archaeal genomes based on DNA structural parameters. *MicrobiologyOpen*, 10, e1230. https://doi.org/10.1002/mbo3.1230

APPENDIX 1

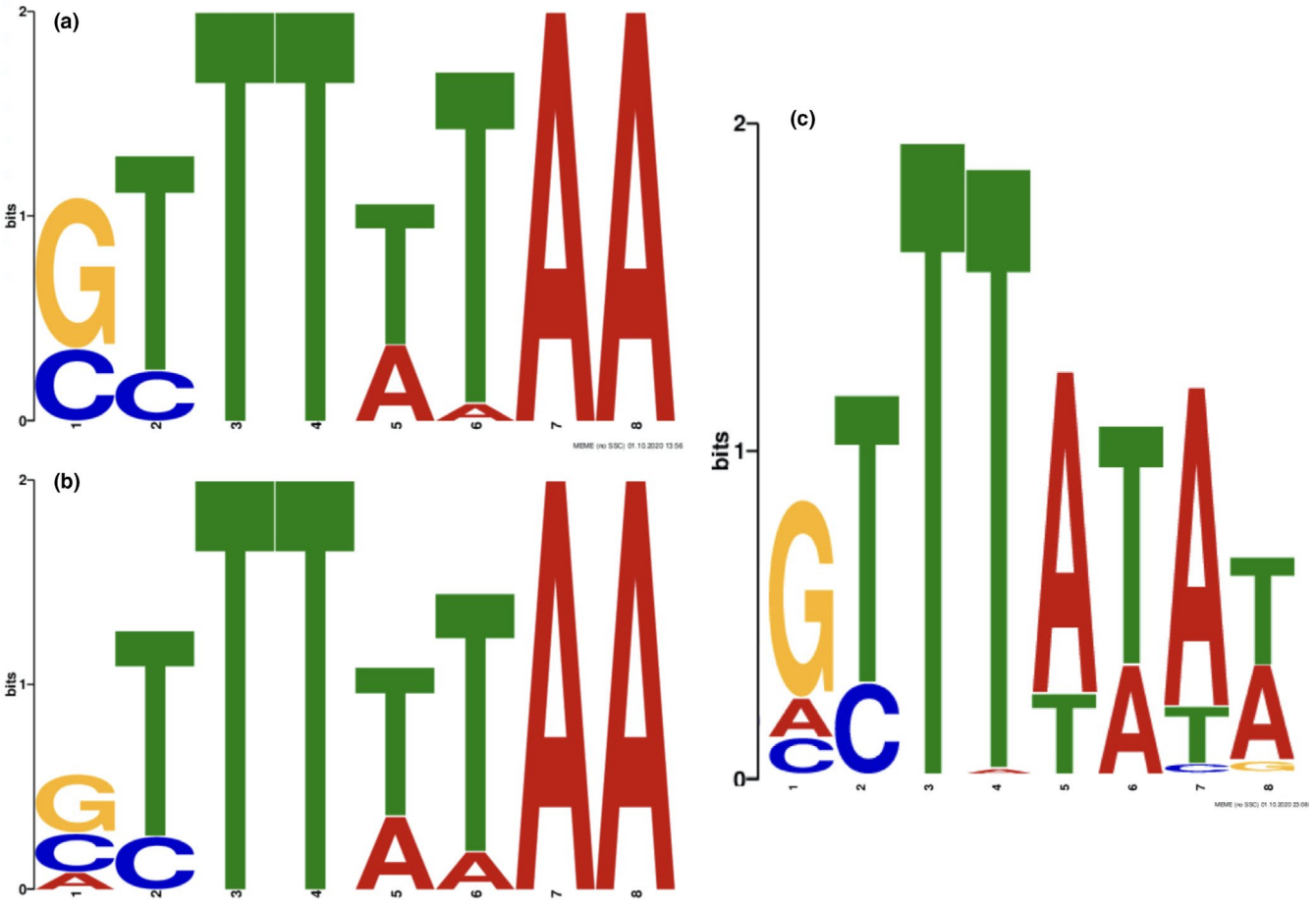

Figure A1 *H. volcanii*, *T. kodakarensis*, and *S. solfataricus* TATA motifs identified by the MEME suite. S1 (a) indicates *H. volcanii*, from which the resulting motif was found in a median position downstream of the TSS of 31 bps. The e-value of this motif is 1.8 e-092; it has been found in 382 sites; its relative entropy is 12.1 and; information content =13.2. S1 (b) indicates *T. kodakarensis*, motifs located in a median distance of 30 bps downstream of the TSS. Its e-value is 1.3 e-017; this motif has been located in 257 sites; relative entropy and information content 12.3 and 12.4, respectively. S1 (c) represents the TATA motif found in *S. solfataricus* found in a median distance of 30 bps downstream the TSS. The e-value =6.1e-022, site count 192, relative entropy =12.5 and, information content =16.2

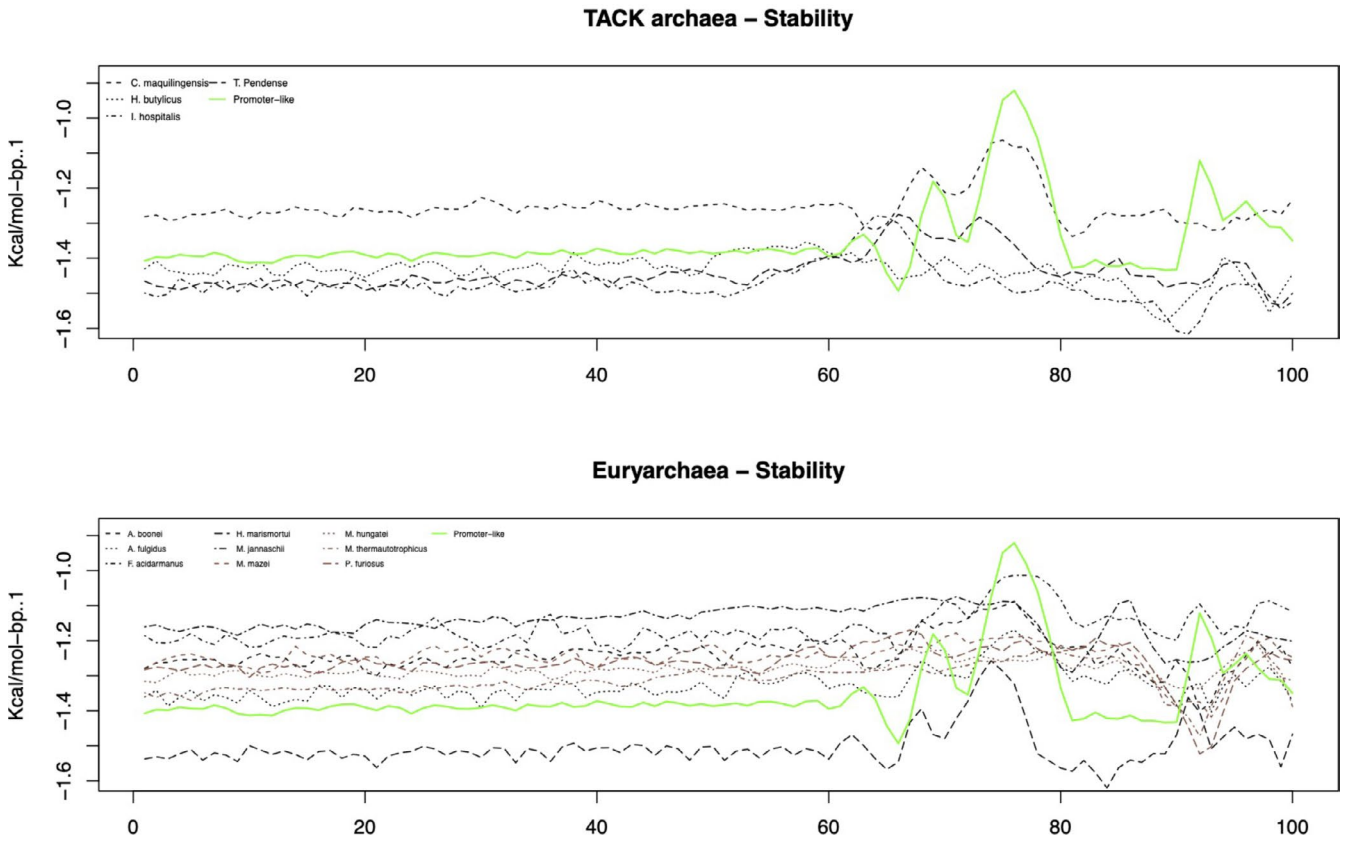

Figure A2 DNA Duplex Stability signal comparison of promoters and upstream regions of thirteen other archaea. The green line (promoter-like) represents the average formed upon experimentally validated promoter of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis* to be compared with upstream sequences of thirteen other archaea divided into two phylogenetic families

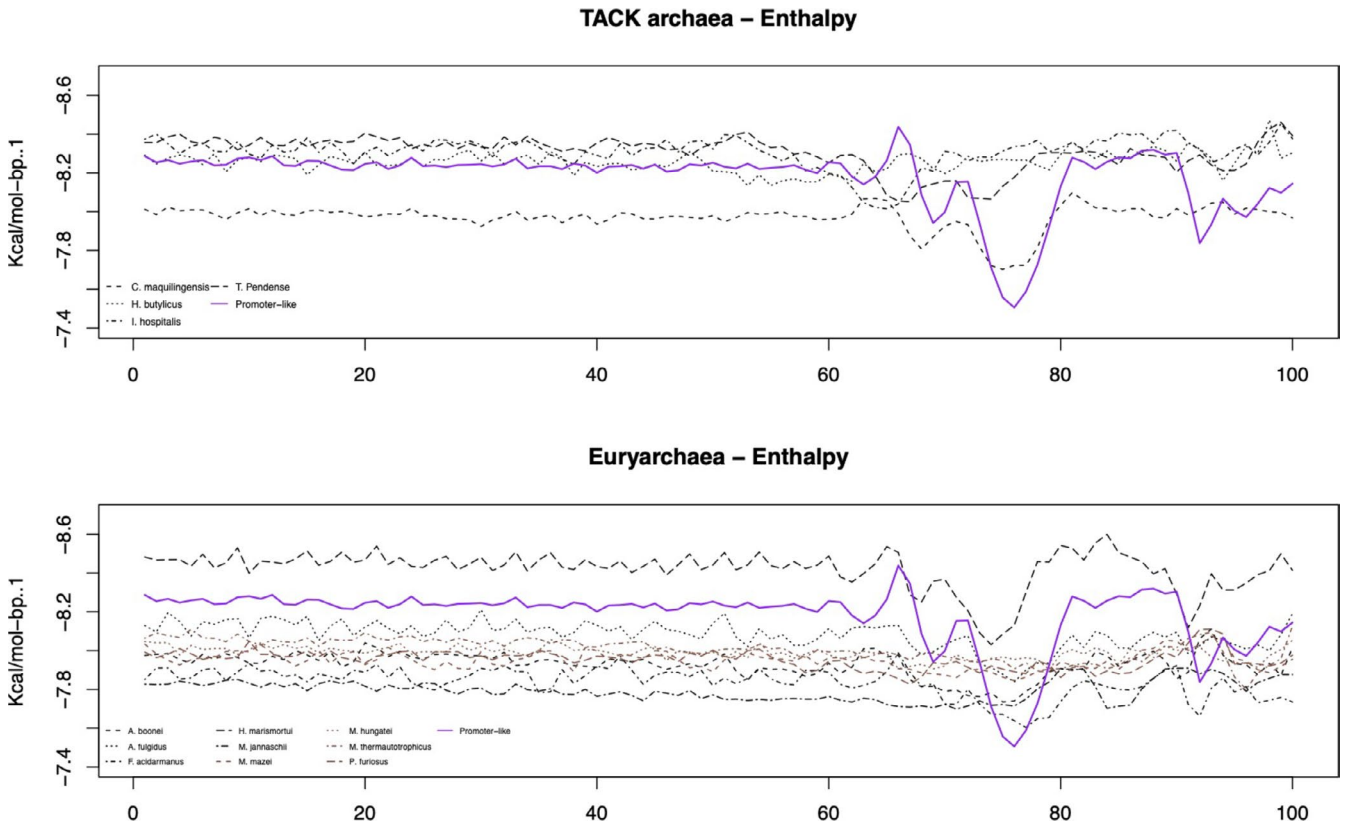

Figure A3 Enthalpy signal comparison of promoters and upstream regions of thirteen other archaea. The purple line (promoter-like) represents the average formed upon experimentally validated promoter of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis* to be compared with upstream sequences of thirteen other archaea divided into two phylogenetic families

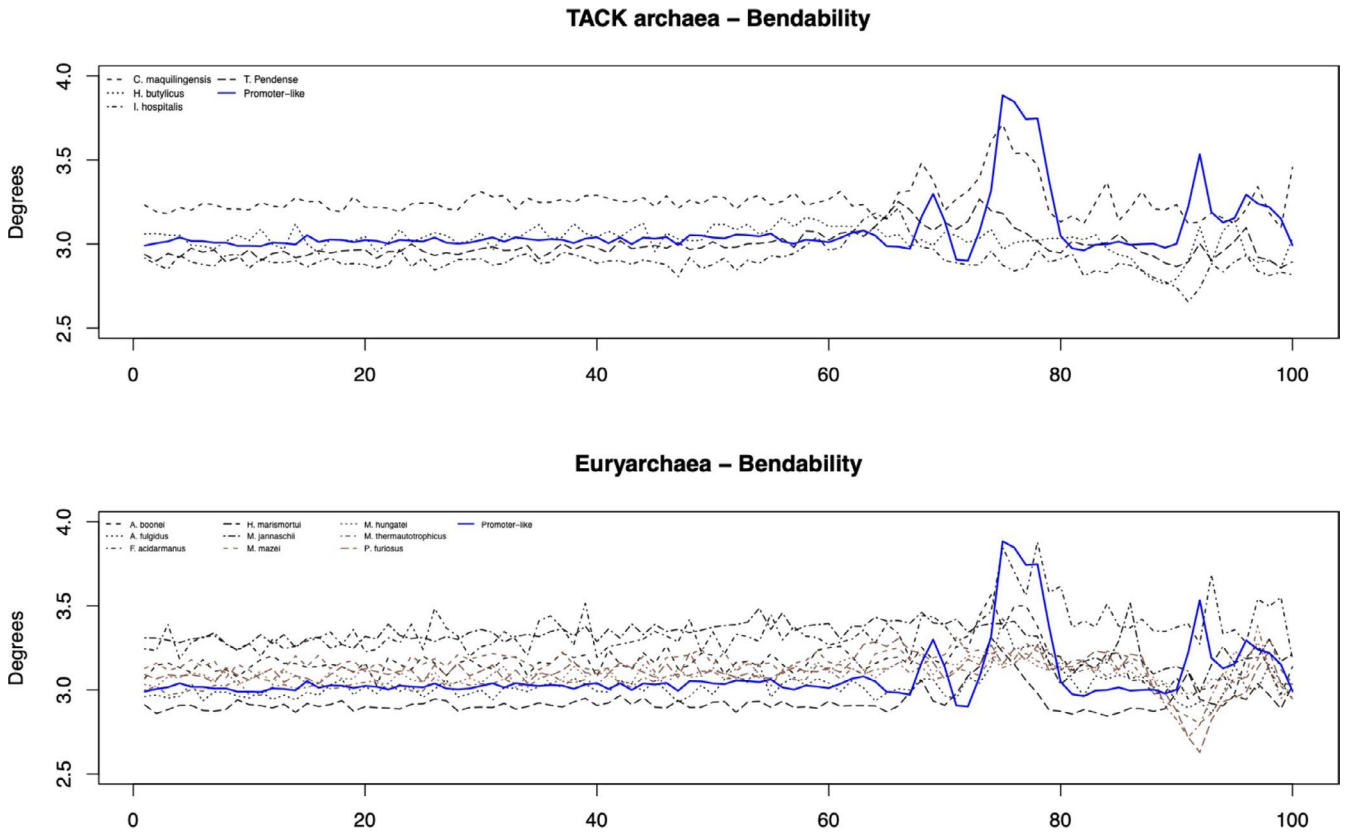

Figure A4 Bendability signal comparison of promoters and upstream regions of thirteen other archaea. The blue line (promoter-like) represents the average formed upon experimentally validated promoter of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis* to be compared with upstream sequences of thirteen other archaea divided into two phylogenetic families

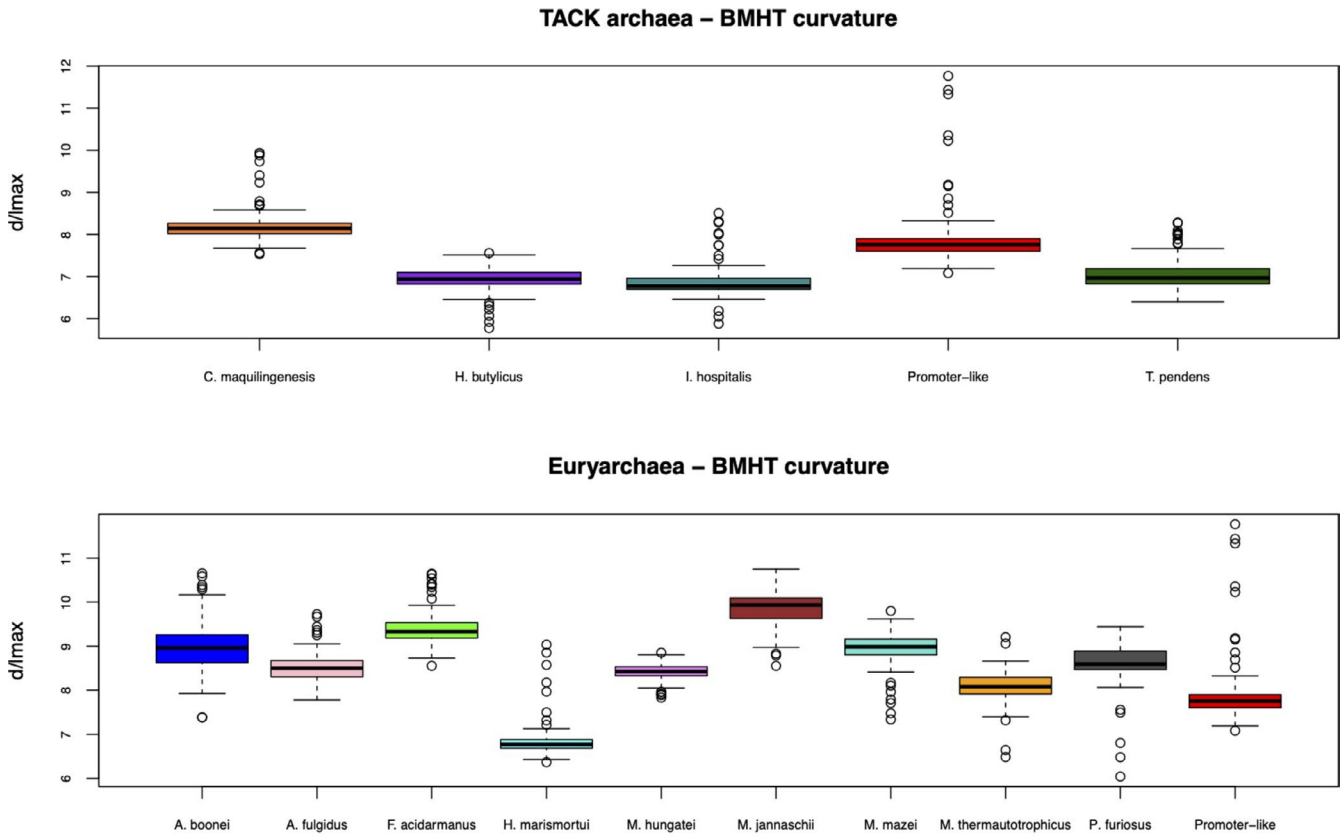

Figure A5 Boxplots of promoters and upstream regions of thirteen other archaea converted to BMHT curvature. The boxplots represent statistical comparisons between the promoter-like profile, (red), formed upon experimental data of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis*. The $p < 2e\text{-}16$ values obtained by the nonparametric Kruskal–Wallis test conveyed statistical significance in the averages of both groups

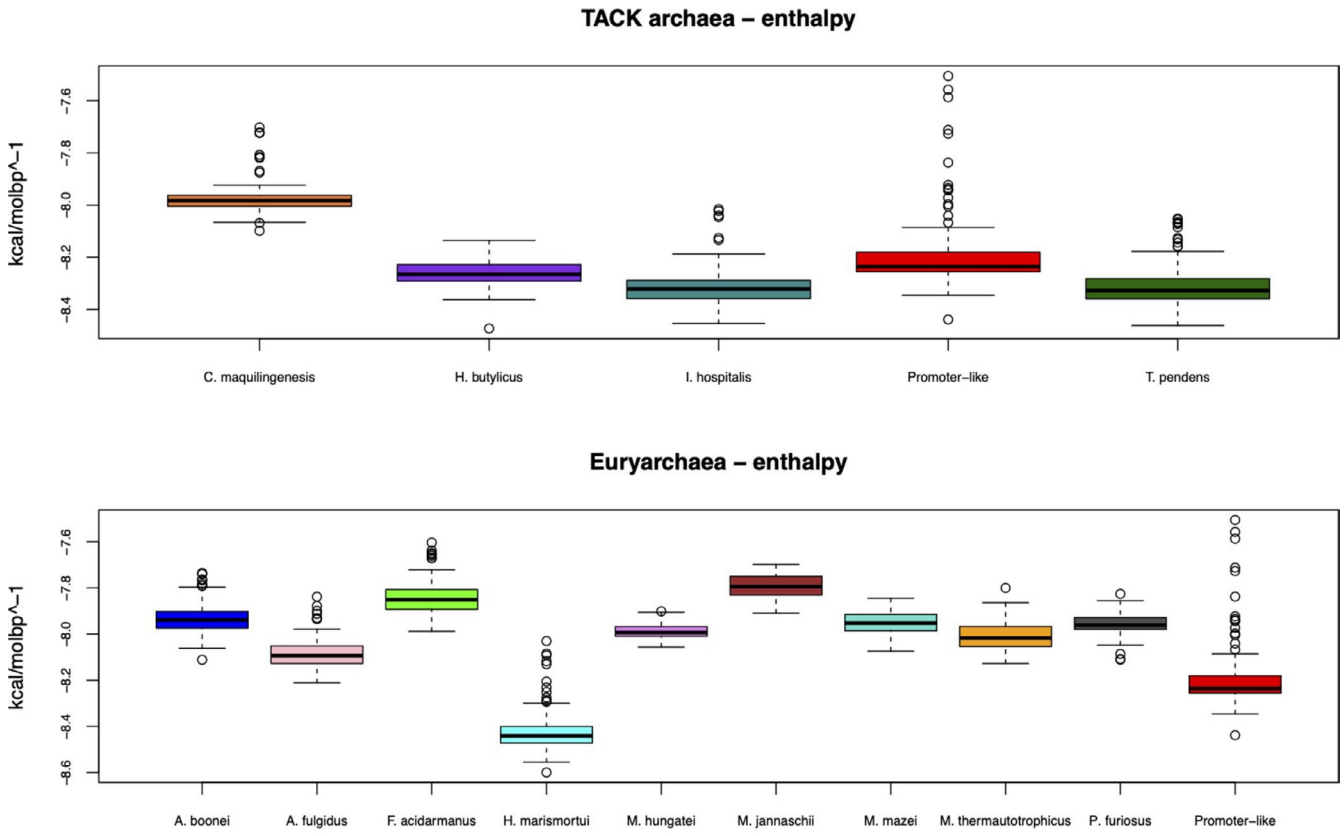

Figure A6 Boxplots of promoters and upstream regions of thirteen other archaea converted to enthalpy. The boxplots represent statistical comparisons between the promoter-like profile, (red), formed upon experimental data of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis*. The $p < 2e\text{-}16$ values obtained by the nonparametric Kruskal–Wallis test conveyed statistical significance in the averages of both groups

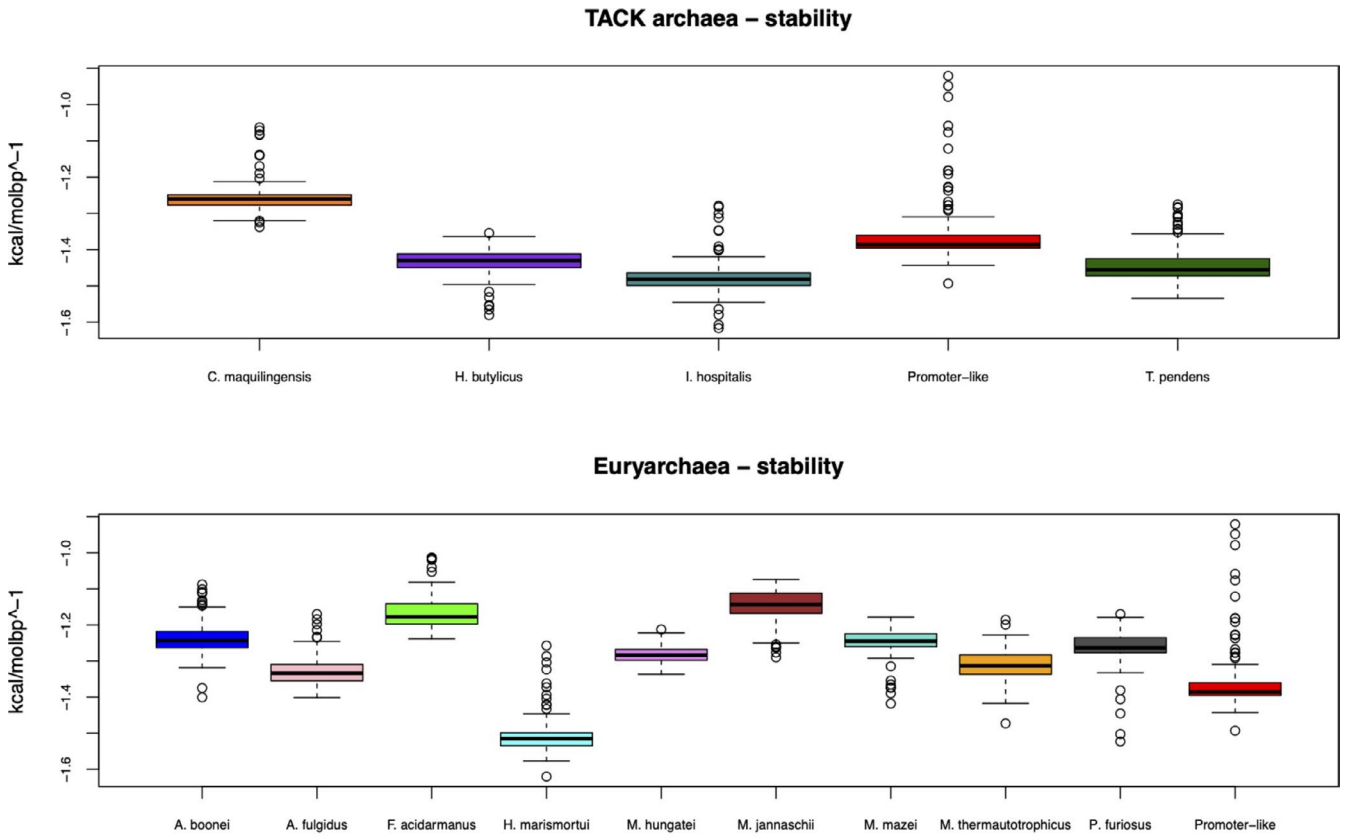

Figure A7 Boxplots of promoters and upstream regions of thirteen other archaea converted to stability. The boxplots represent statistical comparisons between the promoter-like profile, (red), formed upon experimental data of *H. volcanii*, *S. solfataricus*, and *T. kodakarensis*. The $p < 2e\text{-}16$ values obtained by the nonparametric Kruskal–Wallis test conveyed statistical significance in the averages of both groups