DNA Research **19**, 103–115, (2012) doi:10.1093/dnares/dsr045
Advance Access Publication on 14 February 2012

Unveiling *Mycoplasma hyopneumoniae* Promoters: Sequence Definition and Genomic Distribution

Shana de Souto Weber[1], Fernando Hayashi Sant'Anna[1], and Irene Silveira Schrank[1,2,*]

Centro de Biotecnologia, Programa de Pós-graduação em Biologia Celular e Molecular, Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves 9500, C.P. 15005, CEP 91501-970, Porto Alegre, RS, Brazil[1] and Departamento de Biologia Molecular e Biotecnologia - Centro de Biotecnologia, Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves 9500, C.P. 15005, CEP 91501-970, Porto Alegre, RS, Brazil[2]

*To whom correspondence should be addressed. Tel. +55 51-33086055. Fax. +55 51-33087309.
E-mail: irene@cbiot.ufrgs.br

Edited by Katsumi Isono
(Received 24 July 2011; accepted 10 December 2011)

Abstract

Several *Mycoplasma* species have had their genome completely sequenced, including four strains of the swine pathogen *Mycoplasma hyopneumoniae*. Nevertheless, little is known about the nucleotide sequences that control transcriptional initiation in these microorganisms. Therefore, with the objective of investigating the promoter sequences of *M. hyopneumoniae*, 23 transcriptional start sites (TSSs) of distinct genes were mapped. A pattern that resembles the σ^{70} promoter −10 element was found upstream of the TSSs. However, no −35 element was distinguished. Instead, an AT-rich periodic signal was identified. About half of the experimentally defined promoters contained the motif 5′-TRTGn-3′, which was identical to the −16 element usually found in Gram-positive bacteria. The defined promoters were utilized to build position-specific scoring matrices in order to scan putative promoters upstream of all coding sequences (CDSs) in the *M. hyopneumoniae* genome. Two hundred and one signals were found associated with 169 CDSs. Most of these sequences were located within 100 nucleotides of the start codons. This study has shown that the number of promoter-like sequences in the *M. hyopneumoniae* genome is more frequent than expected by chance, indicating that most of the sequences detected are probably biologically functional.

Key words: *Mycoplasma*; promoter; transcription; sigma; matrix

1. Introduction

The genus *Mycoplasma*, composed of bacteria that have no cell wall and have extremely reduced genomes, includes several species of medical or veterinary significance. *Mycoplasma hyopneumoniae* is an important swine pathogen, causing worldwide economic losses in the livestock industry.[1] In recent years, many *Mycoplasma* species have had their genomes completely sequenced, including four strains of *M. hyopneumoniae*.[2–4] Their genomes are ∼900 kb in length and contain ∼700 genes.

The analysis of genomic data shows that *Mycoplasma* genomes contain a small number of genes related to transcription. In the Clusters of Ortholog Groups (COG) classification, there are 20 genes implicated in this process in *M. hyopneumoniae* strain 7448, corresponding to ∼3% of the total coding sequences (http://www.ncbi.nlm.nih.gov/sutils/coxik.cgi?gi=18652). Comparatively, 353 transcription-related genes are found in *Bacillus subtilis*, accounting for 7.4% of the total CDSs (http://www.ncbi.nlm.nih.gov/sutils/coxik.cgi?gi=27).

Like other *Mycoplasma* species, *M. hyopneumoniae* lacks many regulatory elements, including two-component systems and the transcription termination factor Rho.[5] Furthermore, only a single σ factor has been identified in all the *Mycoplasma* genomes

analysed, while *Escherichia coli* has at least six σ factors[6] and *B. subtilis* has at least 18.[7] These observations suggest that mycoplasmas have transcriptional regulatory mechanisms that are unique among bacterial species.

The identification of promoter sequences is an important step towards understanding gene regulation; however, there are few studies about the nucleotide sequences that control transcriptional initiation in *Mycoplasma*. A fundamental study was published more than 10 years ago by Weiner *et al.*,[8] in which several putative *Mycoplasma pneumoniae* promoters were identified by primer extension coupled with analysis using *E. coli* σ^{70} matrices. The defined sequences were used to derive an improved matrix for promoter prediction in this species.

In *M. hyopneumoniae*, very few promoters or transcriptional start sites (TSSs) have been determined. Therefore, with the goal of investigating *M. hyopneumoniae* promoters, 23 gene TSSs were mapped, and their adjacent upstream regions were examined for overrepresented sequences. The data gathered were then used to build species-specific position-specific scoring matrices (PSSMs), which were further evaluated in relation to their predictive performance. The best PSSM was utilized to scan for putative promoters upstream of all coding sequences of the *M. hyopneumoniae* genome.

2. Materials and methods

2.1. Bacterial strains and culture conditions

Mycoplasma hyopneumoniae strain 7448 was cultured in 15-ml Falcon tubes containing 5 ml of Friis medium[9] at 37°C for ~48 h with gentle agitation in a roller drum. *Escherichia coli* XL1-Blue was cultured at 37°C in Luria-Bertani (LB) medium, which was supplemented with 100 μg/ml of ampicillin when required. For blue/white colony selection, 40 μg/ml of X-gal (5-bromo-4-chloro-3-indolyl-β-D-galactopyranoside) and 0.3 mM isopropyl-β-D-thiogalactopyranoside were added to the LB agar.[10]

2.2. DNA manipulations, oligonucleotides and sequence analysis

DNA purifications from agarose gel bands were performed with the NucleoSpin® Extract II kit (Macherey-Nagel GmbH & Co. KG, Düren, Germany) according to the manufacturer's instructions. The *Sma*I-digested plasmid pUC18 was utilized in cloning procedures. DNA ligation, transformation by electroporation, colony polymerase chain reaction (PCR), plasmid extraction and agarose gel electrophoresis were performed using standard methods.[10] The 5′ RACE adapter, the primers 5′ rapid amplification of cDNA ends (RACE) Outer and 5′ RACE Inner were provided in the First Choice RNA ligase-mediated (RLM)-RACE kit (Ambion, Inc., Austin, TX, USA). Gene-specific primers employed in the 5′ RLM-RACE analysis are listed in Supplementary Table S1. The primers M13 forward and M13 reverse (Invitrogen™, Carlsbad, CA, USA) were utilized in the screening of clones and in the sequencing reactions. Sequencing was performed using the Dyenamic *et* dye terminator cycle sequencing kit (Healthcare, Waukesha, WI, USA) and a MegaBACE 1000 DNA Analysis System automated sequencer (Healthcare).

2.3. RNA isolation

Total RNA was isolated from a 25 ml culture of *M. hyopneumoniae* strain 7448. Cells were harvested by centrifugation at $3360 \times g$ for 15 min and resuspended in 1 ml of TRIzol (Invitrogen). The cell suspension was then processed according to the manufacturer's protocol. Subsequently, 50 μg of RNA was treated with RQ1 RNase-Free DNase (Promega Corporation, Madison, WI, USA), followed by purification and concentration with the NucleoSpin® RNA Clean-up XS kit (Macherey-Nagel GmbH & Co. KG).

2.4. 5′ RLM-RACE

To identify TSSs, the strategy described by Bensing *et al.*[11] was employed. This methodology was performed using the First Choice RLM-RACE kit (Ambion, Inc.) following the manufacturer's protocol, except that the calf intestinal phosphatase treatment was not carried out. Briefly, a 16 μl of reaction mixture containing 10 μg of DNA-free RNA, tobacco acid pyrophosphatase (TAP) buffer and 20 U RNase Inhibitor (Fermentas) was divided into two aliquots, one of which received 2 μl of TAP enzyme (TAP+ reaction) and the other an equal volume of water (TAP− reaction). After TAP treatment, both samples were processed identically in the 5′ RACE adapter ligation and reverse transcription steps. Once cDNA was obtained, three nested PCRs were carried out for each gene: TAP+, TAP− and the negative control. All the reactions were performed in a total volume of 25 μl containing 1.25 mM $MgCl_2$, 1× Taq buffer, 0.02 mM of each deoxynucleotide triphosphate (dNTP), 1 U Taq DNA polymerase (Ludwig Biotec, Porto Alegre, Brazil), 10 pmol of the gene- and adaptor-specific primers and 0.5 μl of the template. The outer 5′ RLM-RACE PCR was done with cDNA as the template, the 5′ RACE outer primer and the gene-specific outer primer. The inner 5′ RLM-RACE PCR was done using an aliquot of the outer 5′ RLM-RACE PCR as the template, the 5′ RACE inner primer and the gene-specific inner primer. Amplifications were performed using the touchdown technique, and the products were analysed in 1.2−2% agarose

gels. Differential DNA gel bands present in the TAP-treated samples (fragments derived from unprocessed RNA), but not in the TAP-untreated samples, were purified and cloned. Clones were screened by colony PCR for the presence of the insert and then sequenced.

2.5. Sequence logos

Sequence logos were created using the WebLogo site (http://weblogo.berkeley.edu/).[12,13] Experimentally defined σ^{70} promoter sequences of different bacteria were utilized, relying on the alignments proposed by the respective authors (Supplementary Table S2). The following numbers of promoter sequences were used to generate the logos: 25 sites of *Sinorhizobium meliloti*,[14] 59 sites of *E. coli*,[15] 142 sites of *B. subtilis*,[16] 41 sites of *Chlamydia trachomatis*,[17] 35 sites of *M. pneumoniae*,[8] 25 sites of *Prochlorococcus marinus*,[18] 21 sites of *Campylobacter jejuni*[19] and 23 sites of *M. hyopneumoniae*. Genome size and G + C content were obtained from the genomes deposited in the National Center of Biotechnology Information database (www.ncbi.nlm.nih.gov/).

2.6. PSSMs construction

The 5′ regions of the TSSs determined by RLM-RACE were examined for sequence patterns using the Local-Word-Analysis tool[20] from Regulatory Sequence Analysis Tools[21,22] (RSAT) (http://rsat.ulb.ac.be/). The first 50 bases upstream of the TSSs were analysed, searching for motifs composed of six or four nucleotides, applying a window with a fixed width of 10 nts (for motifs of 6 nts) and a fixed width of 5 nts (for motifs of 4 nts), and a background model that considered all upstream regions of the *M. hyopneumoniae* strain 7448 genes, preventing overlap with upstream open reading frames (ORFs). Overrepresented motifs located four to eight bases upstream of the TSS were manually aligned with BioEdit 7.0,[23] and this alignment was used to build a weight matrix of 12 columns. In addition, other two matrices of 14 and 16 columns were derived using the matrix-building programs MEME[24] and Wconsensus[25] (http://ural.wustl.edu/consensus/), respectively. For building these matrices, 25 bases upstream of the TSSs were analysed with an undefined motif width and the Bernoulli model as the background. In order to mitigate the overfitting problem, the matrices were rebuilt eliminating repeated sites.

2.7. Data set

All analyses were carried out with the sequences obtained from the complete genome of *M. hyopneumoniae* strain 7448, available at NCBI under the accession code NC_007332. The data sets used for both Matrix-Quality and Matrix-Scan procedures were extracted from all the *M. hyopneumoniae* protein-coding genes using the Retrieve Sequence tool from RSAT. The 657 extracted sequences consisted of up to 250 bases upstream (without overlap with the upstream open reading frame) and 50 bases downstream of the annotated start.

2.8. PSSMs performance evaluation

The ability of each of the three PSSMs to discover functional binding sites in the data set sequences was evaluated using the Matrix-Quality[26] program from RSAT.

The following parameters were applied: one pseudo-count was used for correction of the matrix; pseudo-frequencies were set at 0.01. As background, Markov orders from 0 to 4 were tested using the whole set of upstream noncoding sequences of the *M. hyopneumoniae* strain 7448 genome. Comparative analyses of the normalized weight distribution (NWD) curves, obtained from Matrix-Quality, were carried out to decide which matrix and Markov order to use. The trade-off between the estimation of the false-positive rate (FPR) and the sensitivity of the matrix was assessed using receiver-operating characteristic (ROC) curves, containing a leave-one-out (LOO) evaluation of the positive set (sequences used to build the matrix). Finally, as an additional negative control, the empirical and theoretical distributions of the original matrix were compared with the average of 10 column-permuted PSSMs, which were obtained with the Permute-Matrix tool from RSAT.

2.9. Prediction of promoters

The putative *M. hyopneumoniae* promoters were identified using the 12-column weight matrix on the sequence data set through the Matrix-Scan program.[27] The parameters were set as in Matrix-Quality, except that the Markov order was set at 1. The score threshold was determined by comparing the score distribution between the predicted promoters from the sequence data set (correct orientation) with those found in the reverse complement of the sequence data set (incorrect orientation). This analysis excluded intergenic sequences present between genes that are transcribed in divergent directions. The score that resulted in a considerable reduction in putative promoters in the incorrect orientation was selected as the threshold value.

3. Results

3.1. Mapping of TSSs

Initially, the genes for the study of *M. hyopneumoniae* promoters were chosen based on two criteria:

(i) genes annotated as hypothetical were excluded, since it was not known whether they were transcribed and (ii) genes chosen had a divergent upstream gene, thus ensuring that they did not lie inside an operon, and that, consequently, there was a promoter immediately upstream of them. About a quarter of the 79 genes that met these criteria were selected (Supplementary Table S3). The mapping of the TSSs was performed using the 5′ RLM-RACE technique, which allows distinction between primary and processed transcripts on the basis of the phosphorylation state of their 5′ ends. In this process, based on the comparison of 5′ RLM-RACE products derived from RNA treated with TAP and from untreated RNA, it is possible to identify full-length transcripts, since TAP-treated samples include both primary and processed transcripts, while untreated samples include only the processed ones. Thus, the amplification products from TAP-treated RNA samples contained a specific or at least an enhanced signal from primary transcripts compared with untreated RNA samples. Amplification products derived from the 5′ ends of intact transcripts were cloned and sequenced (Supplementary Table S4).

The analysis of 10 or more independent clones for each gene revealed that, in many cases, the 5′ end of the transcripts varied by a few nucleotides in length. In general, the longest sequence was the most common among the clones sequenced. One or two shorter sequences, differing by no more than six nucleotides, were also relatively frequent in eight genes (Supplementary Table S4). These could represent alternative TSSs or could have originated from processed transcripts that were co-purified with the primary ones, since both are present in the TAP-treated samples and may have small length differences. Given the latter assumption, the 5′ nucleotide of the largest sequence of each gene was considered to be the TSS.

Five genes (*sips*, P97-like, *pgk*, *pyrH* and *ktrA*) had additional nucleotides at the 5′ end of their transcripts that were not expected from the genomic sequence (data not shown). The extra nucleotides consisted of one to six adenosines within a homopolymeric region composed of at least three adenosines. In these cases, the last 5′ templated nucleotide was considered to be the TSS.

Overall, the TSSs for 23 *M. hyopneumoniae* genes were identified (Table 1). Four TSSs were found inside of their respective genes: 34 bp within *licA*, 14 bp within *gyrA* and 1 bp within MHP7448_0279 and *dam*. In these cases, the next in-frame start codon downstream of the TSS was assumed to be the true start codon. The distances between TSSs and the gene starts ranged from 143 bp in MHP7448_0360 to 1 bp in *ktrA*. The genes *rplJ* and MHP7448_0198 also had distant TSSs, 100 and 137 bp from their start codons, respectively, while the TSSs of *licA*, *glyA*, MHP7448_0279 and *leuS* were situated <10 bp from their start codons. Further analysis found that 80% of the transcripts initiated with an adenosine residue.

3.2. *Identification of promoter elements*

The 23 experimentally determined TSSs were aligned and the sequences immediately 5′ to them were examined for nucleotide patterns that could comprise promoter elements. The occurrence of locally overrepresented sequences was detected using the Local-Word-Analysis tool. When looking for motifs of six nucleotides, 21 of the 23 genes had the patterns TATAAT or TAAAAT within 5−8 nts of the TSS (Table 1). Additional variants were found in the remaining two genes with multiple em for motif elicitation (MEME) and Wconsensus, which recognized the motifs AAAAAT and TACAAT in the *recA* and *ktrA* genes, respectively (Table 1). Four nucleotide positions of these hexamers were invariant. However, thymidine was the first base in 22 (96%) and the third base in 16 (70%) of them. Therefore, the consensus sequence was TATAAT, which is identical to the canonical σ^{70} promoter −10 element.[15]

The alignment of the sequences using the −10 hexamers revealed additional conserved elements. The base immediately 3′ of the −10 hexamer was thymine in 73% of the sequences (Table 1). Moreover, there was considerable conservation in the bases upstream of the −10 element. The Local-Word-Analysis software found the pattern TATG in eight of the genes, one nucleotide upstream of the −10 element (Table 1). This motif matches the consensus 5′-TRTGn-3′, an extended −10 region commonly found in Gram-positive bacteria that is also known as the −16 element.[28,29] In addition, the dinucleotide TG (the major determinant of −10 extended elements) was found one base upstream of the −10 hexamer in another three genes, so 11 (48%) promoter sequences contained a probable extended −10 element.

While it was possible identify the putative −10 and −16 elements, no conserved pattern corresponding to a −35 element was found (Table 1). Instead, a periodic AT-rich sequence was seen when a sequence logo was created (Fig. 1).

3.3. *Comparison with other σ^{70} bacterial consensus sequences*

Mycoplasma hyopneumoniae promoter sequences were compared with other σ^{70} promoters from different microorganisms. The alignments of experimentally identified sequences were retrieved to

Table 1. Experimentally defined promoter regions of *M. hyopneumoniae*

Gene		5′-region[a]	−16[b]		−10	TSS[c]	Gap[d]	SC[e]
MHP7448_0026	*sipS*	AAAATCAAAAATTAAAATTGTTTTTT	TATG	A	TAAAAT	ATCAAAG	59	ATT
MHP7448_0039	*recA*	TAAATTTTTCCTTTTTTTATTAAAAT	GTTT	A	AAAAAT	ATTAAATTA	71	TTA
MHP7448_0040	*licA*	TAATTTTTATTTTAAAATTTGAAAAA	TATA	A	TAAAAT	TTCCAGTA	8	ATT[f]
MHP7448_0066	*uvrC*	ACTTCAAGATTTAATTATACCAATTT	TTTG	T	TAAAAT	TATAATA	77	ATG
MHP7448_0101	*clpB*	ACTCTTTACTTTTAAGTGCCAAAAAA	TATG	T	TATAAT	TTATTTTGT	16	TTA
MHP7448_0195	*rpsJ*	TAAAAAAATTATTGAATTTTTATTTT	TTGT	G	TATAAT	TTAATCTTA	68	ATG
MHP7448_0198	P97	ACTTTTTTGTTGCAAAAAAAAAAAAA	AAAA	G	TATAAT	TTTAATTG	137	ATG
MHP7448_0224	*glyA*	TTAAAAAAATTATTTTTTTGTTTTTT	TAGT	G	TATAAT	GTGTAAAA	5	ATG
MHP7448_0225		AATAAAAAATAAAAAAAATTATTTTT	TATG	T	TAAAAT	TATAATCG	87	ATG
MHP7448_0272	P97-like	GGATTTTAGTTACTAAAAAATTAAAA	TATG	G	TATAAT	TTTAATTA	56	ATC
MHP7448_0279		GATTTTTTTTTAAAAATTTTTAAAAA	TAGT	G	TAAAAT	TGTTAAA	2	ATG[f]
MHP7448_0359	*glpK*	AATTTCACAGGGCTCCTTTGGATTAA	AATG	T	TATAAT	TCAAATA	26	ATG
MHP7448_0360	P37	TAAAAAAAATTTATAATTCTTCCTTC	TATG	A	TATAAT	AATTTCA	143	ATG
MHP7448_0427	*efp*	TTCATTTTTAGATTTTTTTTTTTTTT	TATG	C	TATAAT	TTATAGTTA	27	ATG
MHP7448_0490	*pgk*	TGTTTTTTCTTAGTTTTTCAACTTAA	TAGT	T	TATAAT	ATAACA	25	ATG
MHP7448_0513	46K	TCATTTTTTAAAAAAAAATTGATTTT	TATA	G	TATAAT	TTATTTG	35	ATG
MHP7448_0528	*gyrA*	GAAAATTCTTATTAAAACATAAAAAA	TATG	G	TATAAT	TTATACTTA	10	TTA[f]
MHP7448_0535	*pyrH*	TTTTTTAATACATTTTTTTTCAAAAA	TAAA	G	TATAAT	AAAAGA	16	ATG
MHP7448_0545	*ktrA*	GATAATTTTAAAAATTTTTTCAATTT	TGGT	C	TACAAT	TTAGTCA	1	ATG
MHP7448_0619	*rplJ*	AAAAAATACTTTTTTTTATTTTCGCTT	TCTG	G	TATAAT	TCAAAAA	100	TTG
MHP7448_0622	*dam*	TTTTTTAAATAATTTTATCCCTATTT	GCTT	A	TATAAT	TTAGTTTA	14	TTA[f]
MHP7448_0647	*leuS*	ACTTTTGGCTTTTAATTTTAAAAAAT	TATG	C	TATAAT	TTAGGTA	6	ATG
MHP7448_0663	P146	GAGAAATTTTTTTAATTTTTAACTTC	TATA	G	TATAAT	TATTGTA	34	ATG
			••••	•	••••••	• 12-col. PSSM		
			•• ••••	•	••••••	• 14-col. PSSM		
			••• ••••	•	••••••	• 16-col. PSSM		

Black background, nucleotides that occur in more than 80% of the promoters; dark grey background, nucleotides that occur in more than 70% of the promoters; light grey background, guanines that occur in more than 40% of the promoters; dots, positions used in the construction of the different PSSMs.

[a]Note that there was no obvious −35 element (TTGACA) in this region.

[b]Region where the −16 element was found.

[c]TSSs are in bold.

[d]Distance (b) between the TSS and the start codon.

[e]Start codons.

[f]The start codons of the genes *licA*, 0279, *gyrA* and *dam* were redefined, as their TSS were located within the original CDS annotation.

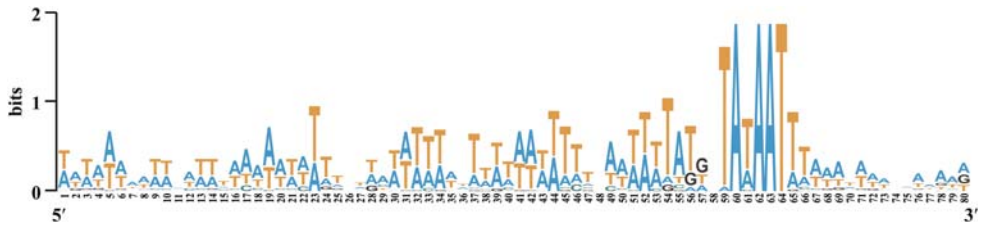


Figure 1. Sequence conservation in the *M. hyopneumoniae* promoter region. Sequence logo derived from the alignment of the 23 defined promoter regions showing the high conservation of the −10 element (positions 59–64), the presence of a semi-conserved −16 element (positions 54–57), the absence of a −35 element and the distinct periodic AT-rich signal extending upstream of the −10 element. The region extending between positions 54 and 65 was used to construct the 12-column PSSM. The vertical axis shows information content in bits. The overall height of the stack indicates the sequence conservation at that position, whereas the height of the nucleotide within the stack indicates its relative frequency at that position.

create logos, which visually represent sequence conservation.

The results presented in Fig. 2 suggest that the occurrence of −35 elements in the σ70 promoter is related to the G + C content of the organism. The promoters of the species *S. meliloti*, *E. coli*, *B. subtilis*, *C. trachomatis* and *M. pneumoniae*, which have a genomic G + C content ≥40%, have the trinucleotide TTG of the −35 element, whereas the promoters of *P. marinus*, *C. jejuni* and *M. hyopneumoniae*, which have a genomic G + C content ≤30.8%, do not have this conserved trinucleotide.

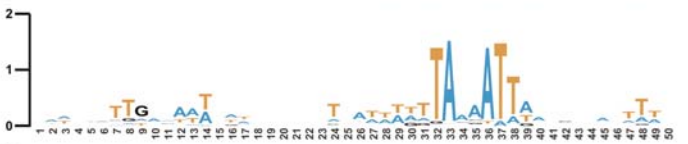
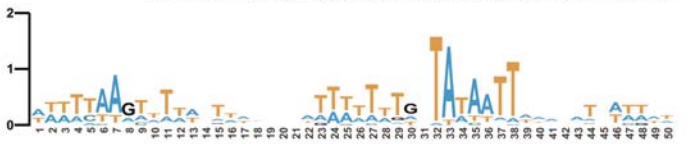
Bacterium	Genome size (kb)	Genomic G+C (%)	Sequence logo (5'–3')
S. meliloti	3,654	62.7	
E. coli	5,498	50.5	
B. subtilis	4,215	43.5	
C. trachomatis	1,042	41.3	
M. pneumoniae	816	40	
P. marinus	1,657	30.8	
C. jejuni	1,616	30.6	
M. hyopneumoniae	920	28.5	

Figure 2. σ^{70}-like recognition sites in different bacterial species. Sequence logos showing the loss of conservation of the −35 signal as the genomic G + C content decreases. The following numbers of promoter sequences were used to generate the logos: 25 sites of *S. meliloti*, 59 sites of *E. coli*, 142 sites of *B. subtilis*, 41 sites of *C. trachomatis*, 35 sites of *M. pneumoniae*, 25 sites of *P. marinus*, 21 sites of *C. jejuni* and 23 sites of *M. hyopneumoniae*. The vertical axis shows information content in bits. The overall height of the stack indicates the sequence conservation at that position, whereas the height of the nucleotide within the stack indicates its relative frequency at that position.

Comparison of these sequence logos also indicated that the −10 element is more conserved in *M. hyopneumoniae* than in the other bacterial species. One noteworthy observation was that this element was preceded by the dinucleotide TG, a feature that is shared with *B. subtilis* and *C. jejuni*, indicating the existence of a −16 element. Another distinct characteristic found was the presence of periodic AT-rich sequences upstream of the −10 elements of *M. hyopneumoniae* and *C. jejuni*.

3.4. *Construction of a PSSM for prediction of* M. hyopneumoniae *promoters*

Manual alignment of the 23 defined *M. hyopneumoniae* promoters was used to create a PSSM of 12 columns (Tables 1 and 2). In order to validate whether this alignment and the positions included in the matrix were appropriate, two other matrices were independently constructed using MEME and the Wconsensus. Both programs included the same 12 positions used in the initial matrix to build their

Table 2. PSSM based on experimentally determined *M. hyopneumoniae* promoters

	−17	−16	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6
A	2	18	3	6	5	1	23	5	23	22	0	5
C	0	2	0	0	3	0	0	1	0	0	0	0
G	1	1	5	11	10	0	0	0	0	0	0	1
T	20	2	15	6	5	22	0	17	0	1	23	17

matrices. However, these latter matrices included a few more positions, generating PSSMs of 14 and 16 columns (Table 1).

Once the three PSSMs were obtained, they were rebuilt, excluding the repeated sites, with the aim of minimizing the problem of overfitting, and then their predictive capacity was assessed and compared in order to choose the best one. Matrix-Quality was used to perform this evaluation. This program relies on a combined analysis of theoretical and empirical score distributions to estimate the capability of a PSSM to distinguish putative binding sites from the genomic background.[26] The theoretical distribution encompasses the matrix scores along a random sequence of infinite length generated using the background model. This indicates the probability of a site scoring above a given weight score by chance, and thus provides an estimate of the FPR.[26] The empirical distribution contains the matrix scores obtained along the sequences of interest (e.g. upstream noncoding sequences), which are composed predominantly of nonbinding sites, interspersed with a few biologically functional sites.[26] Both distributions were calculated using the three PSSMs. For the empirical distribution, the sequence set comprised up to 250 bases upstream and 50 bases downstream of the start codon from all *M. hyopneumoniae* protein-coding genes (downstream bases were also scanned because some TSSs were found within genes). As a background model, the whole set of the upstream noncoding sequences of the *M. hyopneumoniae* genome was used, testing different Markov orders (0–4), since this affects the weight score computation and, consequently, the performance of the matrices. The discriminatory capability of each matrix coupled with each Markov order was assessed by comparison of the empirical and theoretical score distributions.

The difference between the two distributions indicates the discriminative power of the matrix, which can be expressed by computation of the NWD curves.[26] In this analysis, the weight score difference (WD) between the weight scores observed in empirical and theoretical distributions is calculated at each frequency value. As larger matrices allow higher scores, the WD is divided by the number of matrix columns to obtain the NWD, which allows that matrices of different lengths to be compared. All matrices performed better using a Markov order of 1 (Supplementary Fig. S1). Comparison of these matrices using this background model showed that the matrix of 12 columns yielded the highest NWD values (Fig. 3), indicating that this PSSM was the best one to discriminate putative promoter sequences from the noncoding genomic background.

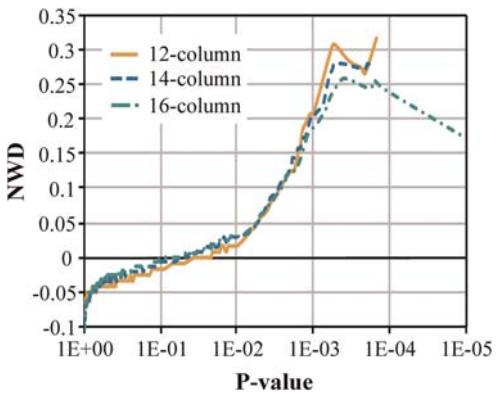


Figure 3. Performances of the 12-, 14- and 16-column PSSMs using a Markov order of 1 as the background model. Each curve shows the normalized weight score difference (NWD) calculated from theoretical and empirical distributions obtained for each matrix using a Markov order of 1. The higher the NWD value, the better the matrix distinguished putative sites from the noncoding genomic background.

Once the PSSM and the background model were defined, additional analyses were performed. In order to generate a complementary negative control, the same data set used for the empirical distribution was scanned using column-permuted matrices derived from the 12-column PSSM. Figure 4 shows that the mean of the score distributions of 10 permuted matrices overlapped the theoretical distribution. This confirmed that the theoretical distribution can be considered an appropriate estimate of the FPR, and that the divergence observed in the original PSSM distribution corresponded to sites specifically detected by this matrix in the genome.[26]

3.5. *Score threshold determination*

The curves of the theoretical and empirical distributions of the 12-column PSSM began to separate from each other around a weight score of 3 (Fig. 4), which is probably indicative of the presence of functional

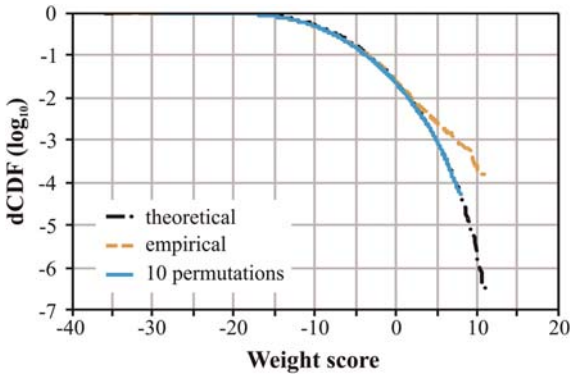


Figure 4. Weight score distributions for the 12-column PSSM. The curves of the theoretical (black; dashed-dotted line) and empirical (orange; dashed line) distributions obtained with the 12-column PSSM began to separate at a weight score of 3, indicating that the promoters were being distinguished from the genomic background. Note that the mean of the score distributions of 10 column-permuted matrices (blue; solid line) overlaps with the theoretical distribution, confirming that the theoretical distribution can be considered an appropriate estimation of the FPR. The theoretical score distribution was estimated with a Markov model of order 1 using the whole set of upstream noncoding sequences of *M. hyopneumoniae*. The empirical score distribution was obtained with a sequence set composed of the 250 bases upstream and 50 bases downstream of the start codon of all *M. hyopneumoniae* protein-coding genes. The dCDF (ordinate) indicates the probability of observing a site scoring higher than or equal to a given weight score (abscissa).

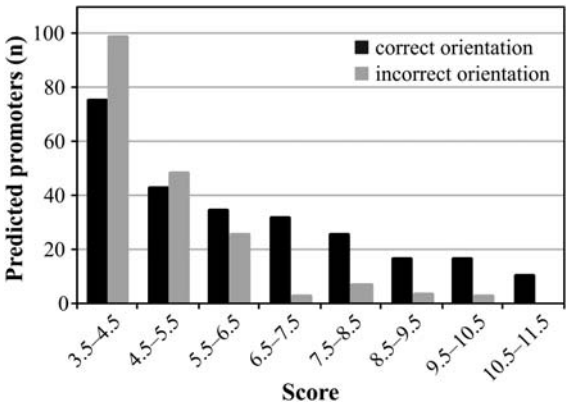


Figure 5. Weight score threshold definition. Distribution of the scores of the correctly and incorrectly oriented promoters predicted in the *M. hyopneumoniae* intergenic regions. Note that from score 6.5, the frequencies of incorrectly oriented promoters are much smaller than the frequencies of correctly oriented promoters.

binding sites. At this score value, the decreasing cumulative distribution function (dCDF, indicates the *P*-value, i.e. the probability to obtain by chance a weight score higher than or equal to a given value) in theoretical distribution is 4.1×10^{-3} (3.86×10^{-3} in the permuted matrix distribution) and in the empirical distribution is 5.8×10^{-3}. It means that for ~6 sites found in the upstream gene sequences, one could expect that ~4 of those were false-positives. Hence, the incidence of false-positives in relation to the observed frequency of sites in the target sequences is too high at this point. However, from a score of 3 upwards, the difference between the observed and the expected frequencies gradually increased (Fig. 4). Consequently, the choice of a score threshold that would allow comprehensive promoter identification with a relatively low FPR was necessary.

The threshold score was defined using the complementary approach described by Cases *et al.*[30] This procedure compares the score distributions of predicted promoters that are 'correctly' oriented – that are in the same direction as the downstream gene – with those found in the reverse strand, which are, therefore, 'incorrectly' oriented. The assumption is that false-positives should be homogeneously distributed between both strands, whereas true positives must be correctly oriented.[30]

The target sequences were composed of those from the dataset used for the determination of the empirical distribution, but the sequences located between divergent genes were excluded, as they could have had promoters in both directions. The occurrence of putative promoters in these sequences and their respective reverse complements was determined by Matrix-Scan using the 12-column PSSM and a Markov order of 1 as the background model. The distributions of the correctly and incorrectly oriented promoters are presented in Fig. 5. The incidence of incorrectly oriented promoters considerably diminished with a weight score of 6.5, so this was used as the threshold score for posterior analyses. The estimated FPR at this score was 2.4×10^{-4} (2×10^{-4} in the permuted matrices distribution), whereas the dCDF in the empirical distribution was 1.42×10^{-3}.

The trade-off between the FPR and the sensitivity of the threshold score was assessed using the ROC curve generated by Matrix-Quality analysis (Fig. 6). The sensitivity of a PSSM is the proportion of correct sites detected above the score threshold, and it is estimated by scoring the sites used to build the matrix.[26] This estimation was also performed using the LOO validation, which corrects biases in matrix sensitivity.[26] Figure 6 shows that a FPR of 2.4×10^{-4} (at a score of 6.5) is associated with a sensitivity of 0.65 for the biased curve, and 0.60 for the LOO curve. It is worth noting that the LOO curve and the unbiased curve are not distant from each other, so overfitting was insignificant.

3.6. *Predicted promoters*

After the optimum matrix parameters were defined, the upstream sequences of all 657 *M. hyopneumoniae*

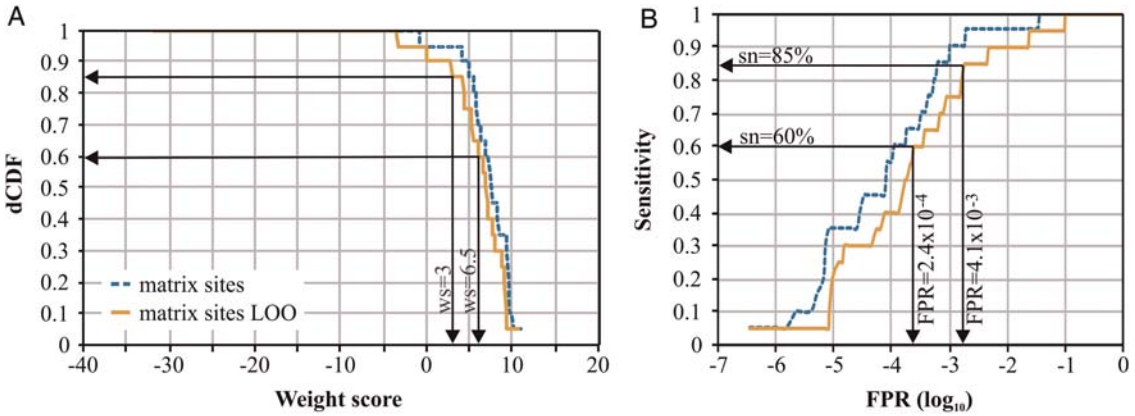


Figure 6. Trade-off between the sensitivity and FPR of the 12-column PSSM. (A) Score distributions of the experimentally defined sites used to build the matrix. Blue (dashed line), biased scores assigned by the matrix to the defined sites. Orange (solid line), unbiased scores obtained using the LOO procedure. The ordinate indicates the probability of observing a site scoring higher than or equal to a given weight score (abscissa). (B) ROC curve indicating the risk of false-positives associated with a specific sensitivity. Both graphs show the difference between the biased (blue; dashed line) and LOO estimations (orange; solid line). The dCDF (ordinate) indicates the sensitivity (fraction of sites detected) and the abscissa shows the corresponding FPR. Note that the dCDF (A) corresponds to the sensitivity (B).

Table 3. *Mycoplasma hyopneumoniae* promoter prediction analysis

	N
Genomic features	
CDSs annotated in the genome	657
CDSs that have an upstream region < 15 bp	201/657 (31%)
CDSs that have a divergent upstream gene	142/657 (22%)
CDSs that have an upstream gene oriented in the same direction	515/657 (78%)
Predicted promoter features (weight score ≥ 6.5)	
Promoters	201
CDSs that have at least one promoter	169/657 (26%)
CDSs that have:	
One promoter	143/169 (84%)
Two promoters	22/169 (13%)
Three promoters	3/169 (2%)
Five promoters	1/169 (<1%)
CDSs that have a divergent upstream gene and have at least one promoter	76/142 (54%)
CDSs that have an upstream gene oriented in the same direction and have at least one promoter	93/515 (18%)
Predicted promoter features (weight score ≥ 4.2)	
Promoters	409
CDSs that have at least one promoter	273/657 (42%)
CDSs that have a divergent upstream gene and have at least one promoter	113/142 (80%)
CDSs that have an upstream gene oriented in the same direction and have at least one promoter	160/515 (31%)

CDSs were scanned for the presence of putative promoters using Matrix-Scan. Table 3 shows the general results of this analysis. Using a threshold score of 6.5, 201 sites were identified upstream of 169 different genes, 26% of the total CDSs.

The vast majority of the CDSs had a single putative promoter, although there were CDSs that had additional sites. In this promoter prediction analysis, 16 of the 23 promoters experimentally mapped scored between 6.9 and 11, six scored between 4.2 and 6.3, and one, the *recA* promoter, did not score above zero. Most of them corresponded to the hit with the highest score, but those of the genes *uvrC*, MHP7448_0198 and *ktrA* were the second best hits (although none of these scored higher than 6.5).

Our analyses detected at least one promoter in 54% of the CDSs that had a divergent upstream gene and in 18% of the CDSs that had an upstream gene oriented in the same direction. However, these proportions were 80 and 31%, respectively, if the threshold score was set at 4.2, the smallest weight score obtained for the experimentally defined promoters.

The distance of the promoters from the start codon was also examined. The majority of the predicted promoters, ~67.5%, were located between 1 and 100 bases upstream of the start codon, with a preponderance located 25–50 bases upstream (Fig. 7). Sixteen promoters were found within the coding sequences of 14 CDSs.

4. Discussion

The transcripts of 23 genes of *M. hyopneumoniae* were analysed in order to map their TSSs. As is usually the case in transcripts of other bacteria,[8,15,16,18] most of those from *M. hyopneumoniae* started with a purine. Our data also showed that many of its gene transcripts had variation at their 5′ ends, suggesting

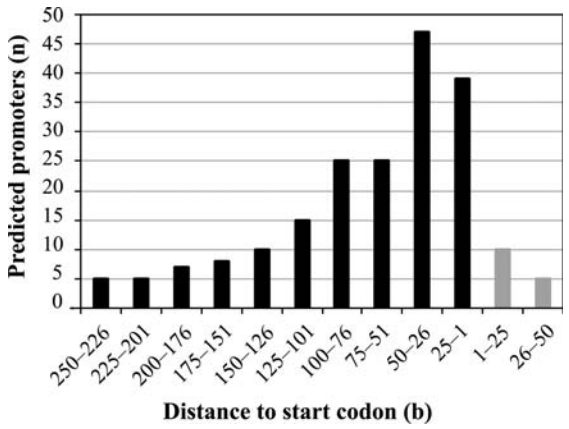


Figure 7. Distance of the predicted promoters from the annotated start codon of the *M. hyopneumoniae* CDSs. Distances were determined for the 201 predicted promoters scoring $\geq$6.5; they were measured from the -10 element to the start codon of the genes. Black bars indicate bases upstream of the start codon, and grey bars indicate bases downstream of the start codon.

the occurrence of heterogeneous TSSs. The heterogeneity observed in some *M. hyopneumoniae* transcripts was due to additional untemplated nucleotides (i.e. nucleotides not expected from the genome sequence), and was probably the result of transcriptional slippage.[31] In this process, the RNA polymerase adds nucleotides, repetitively, to the 3′ end of the nascent transcript, typically within homopolymeric sequences. Differently, the 5′ end of some transcripts had length differences in which the additional nucleotides were identical to the genomic sequence. Such templated heterogeneous 5′ ends have also been seen frequently in the *M. pneumoniae* transcripts.[8]

In addition, a high frequency of transcripts that have just few nucleotides in their 5′ untranslated region, reported in *M. pneumoniae*,[8] was also seen in *M. hyopneumoniae*. Translation can be initiated on the leaderless mRNAs in the three domains of life, but, although they are abundant in Archaea, they are still considered rare in bacteria. Thus, as mentioned by Weiner *et al*.,[8] this high incidence of leaderless transcripts in *Mycoplasma* could be a result of adaptation to a minimal genome with the aim of reducing the genomic space required for initiation of translation.

The only σ factor identified in the *M. hyopneumoniae* genome belongs to the σ^{70} protein family. The σ^{70} factors interact with archetypical promoters that are composed of two main regions: the -35 element (TTGACA) and the -10 element (TATAAT). The upstream regions of experimentally defined TSSs of the *M. hyopneumoniae* genes contained a -10 element, but no obvious -35 element, a structure shared with other low G + C content bacteria.[19] It has been suggested that organisms that have undergone massive reductions in their genome acquired a low G + C content and have also had degradation of their regulatory signals.[32]

Previous studies have demonstrated that transcription can occur when only the -10 element is present, although additional elements, including activator proteins and extended -10 elements (the -16 element), may be involved.[32] Forty-eight per cent of the experimentally characterized *M. hyopneumoniae* promoters contained the -16 element. This proportion is very similar to that found in *B. subtilis*, in which ~45% of promoters possess this element.[33] Studies suggest that the extended elements compensate for the lack of conservation in the -10 and -35 boxes of the promoters.[34] The -16 elements are also found in promoters of other species, including *E. coli* and *C. jejuni*, but are not seen in *M. pneumoniae* promoters.[8,35]

The AT-rich stretches upstream of the -10 element in the promoters of *M. hyopneumoniae* and *C. jejuni* may result in transcriptional enhancement. Petersen *et al*.[36] suggested that they could play a role as specific binding sites or be implicated in DNA curvature. These stretches could also be related to upstream (UP) elements, which can affect promoter recognition and activity.[37] UP elements are AT-rich sequences, typically located in a region from nt -40 to nt -60 (relative to the TSS) that interacts with the C-terminal domain of the α-subunits of RNA polymerase.[37] They have been identified in several bacterial species, and their occurrence increases as the genomic G + C content of the organisms decreases.[38] UP elements can improve the activity of a TGn/-10 promoter in the absence of a good -35 element,[39] and promoters comprising only UP and -10 elements can be recognized by RNA polymerase.[40] Thus, in AT-rich organisms, such as *M. hyopneumoniae*, it is likely that the AT-rich stretches act as UP elements, which may lessen the requirement for -35 hexamers.

While the conservation of the -10 element in both *Mycoplasma* species is evident, the *M. hyopneumoniae* promoters are particularly similar to those of *C. jejuni*. Besides lacking the -35 signal and possessing the extended -10 element, they also have periodic AT-rich stretches upstream of the -10 region. Since *M. hyopneumoniae* (Tenericutes) and *C. jejuni* (Proteobacteria) are phylogenetically distant, their promoter similarities suggest evolutionary convergence, which could be consequence of their high genomic A + T content (~70%).

PSSMs have been widely used to find conserved motifs.[27] A PSSM was defined based on the experimentally defined promoters in order to detect promoter-like sequences along the intergenic regions of the *M. hyopneumoniae* genome.

The promoter scan of *M. hyopneumoniae* sequences found that the pattern detected by the matrix

occurred more frequently than expected, indicating that it did not occur by chance and that it was probably functional in initiating gene transcription. Recent studies based on *E. coli* σ^{70} promoter data were not able to detect these patterns in *Mycoplasma* genomes,[32,41] even suggesting that the existence of promoters in these bacteria was debatable.[41] However, as demonstrated by Weiner *et al*.,[8] the identification of *Mycoplasma* promoters using an *E. coli* matrix is not efficient. Our study has improved on these previous studies by using a species-specific PSSM that accounted for the variability between bacterial species, avoiding biases that might result from using heterologous PSSMs.

Approximately 26% of the CDSs in the *M. hyopneumoniae* genome had at least one identifiable promoter in their upstream region. However, many of the upstream sequences of the CDSs were too short to contain a promoter sequence of 12 nucleotides and a spacer of four nucleotides preceding the TSS. Therefore, the coverage of CDSs that could contain a promoter was greater than estimated. Adams *et al*.[42] have suggested that the upper limits of the intergenic regions in the *M. hyopneumoniae* operons is ~50 bases, and studies have shown that genes that are organized in tandem with intergenic distances much larger than 50 bases can be transcribed in large transcriptional units.[43] These findings indicate that many CDSs are regulated by common promoters, and therefore that not all CDSs necessarily have a promoter in their adjacent upstream regions.

Intergenic regions between divergently oriented genes are the most probable sites to find promoter-like sequences. Our analyses indicated that 54% of the genes with this organization had at least one promoter signal. In contrast, of the 515 genes oriented in tandem, only 93 (18%) had a promoter sequence upstream. The relatively small proportion of in tandem CDSs that possessed promoters was probably attributable to the organization of most of these genes in transcriptional units and, therefore, their transcription might be driven by promoters that are not in the nearest upstream intergenic region.

Although experimental studies have detected the presence of large transcripts in *M. hyopneumoniae*, which could be transcribed from the promoter upstream of the first CDS of the transcriptional unit, our study demonstrates that many internal CDSs may also contain putative promoters. For instance, in the experimentally defined transcriptional unit containing the genes *deoC*, *upp*, MHP7448_0525, *lon* and *tuf*,[44] all the genes, except MHP7448_0525, contain promoter sequences in their upstream regions (with scores varying from 8.4 to 11) (data not shown). This example corroborates the findings of Gardner *et al*.,[43] who demonstrated that, even when transcription does not cease between genes, there is evidence of independent transcriptional initiation by the promoter of the following gene.

Most of the CDSs had a single promoter sequence (84%), but CDSs with multiple promoter sequences were also detected. The *tuf* gene, for example, which is known to be highly expressed, possessed three promoters in its upstream region, two of which overlapped (data not shown). Overlapping signals could promote transcription by recruiting RNA polymerases to the primary promoter sequence.[45] In the absence of a strong promoter, overlapping sites could be non-competitive weak promoters that could produce basal transcription of the downstream genes. On the other hand, they could also negatively regulate transcription through competition between RNA polymerases,[46] or through the induction of a pause in the early steps in elongation.[47,48]

The majority of the putative promoters were found between 1 and 100 bases upstream of the start codon. This is congruent with many previous studies performed in different bacterial species.[36] Some predicted promoter sequences were found within CDSs. This could be because the start codons of these genes were not assigned correctly, or because these putative intragenic signals have an unknown regulatory function.

Although a comprehensive prediction of promoters was performed in this study, many putative signals were not detected using the criteria used for prediction. The main restraint was the threshold score of 6.5. Approximately 30% of the promoters defined experimentally in our study were not detected using this cut-off value. Even the *recA* promoter was not detected using these criteria. The lowest score for the experimentally defined promoters was 4.2; however, at this threshold, about half of the sequences identified were estimated to be false-positives. There are many promoter-like sequences in the genome with scores >4.2 ($\mathrm{dCDF} = 3.46 \times 10^{-3}$), raising the question of how RNA polymerase distinguishes the signals of true promoters from the false-positives. As *M. hyopneumoniae* only has a small number of known regulatory proteins,[2] one might speculate that most of the sequences that score >4.2 are true promoters. Gardner *et al*.[43] found that there is transcription across the majority of the intergenic regions in *M. hyopneumoniae*. However, studies have demonstrated that this species is able to control transcription;[49–53] therefore, the sequence contexts in which the signals are immersed may be a determinant of transcriptional initiation.

In summary, our study has contributed to understanding of transcriptional regulation in *M. hyopneumoniae*, as it has identified basic elements involved in transcriptional initiation and verified their distribution in the upstream regions of protein-coding genes

in this species. Possible applications for the PSSM defined in this study would be refinement of genome annotations and investigation of promoters in closely related species, such as *Mycoplasma hyorhinis* and *Mycoplasma flocculare*.

Acknowledgements: We especially thank Professor Augusto Schrank for valuable suggestions and Alejandra Medina-Rivera for all support with the Matrix-Quality program. We thank Professor Arnaldo Zaha for revising the manuscript. We thank Franciele Maboni Siqueira for supplying her unpublished experimental data. We also thank Bianca Gervini Fávero Bittencourt for the *M. hyopneumoniae* cultures.

Supplementary Data: Supplementary Data are available at www.dnaresearch.oxfordjournals.org.

Funding

This work was supported by grants from the Brazilian National Research Council, the Fundação de Amparo à Pesquisa do Rio Grande do Sul and the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

References

1. Thacker, E.L. 2006, *Diseases for Swine*. In: Straw, B.E., Zimmermann, J.J., D'Allaire, S. and Taylor, D.J. (eds), Iowa State University Press: Ames, pp. 701–17.
2. Minion, F.C., Lefkowitz, E.J., Madsen, M.L., Cleary, B.J., Swartzell, S.M. and Mahairas, G.G. 2004, The genome sequence of *Mycoplasma hyopneumoniae* strain 232, the agent of swine mycoplasmosis, *J. Bacteriol.*, **186**, 7123–33.
3. Vasconcelos, A.T., Ferreira, H.B., Bizarro, C.V., et al. 2005, Swine and poultry pathogens: the complete genome sequences of two strains of *Mycoplasma hyopneumoniae* and a strain of *Mycoplasma synoviae*, *J. Bacteriol.*, **187**, 5568–77.
4. Liu, W., Feng, Z., Fang, L., et al. 2011, Complete genome sequence of *Mycoplasma hyopneumoniae* strain 168, *J. Bacteriol.*, **193**, 1016–7.
5. Fraser, C.M., Gocayne, J.D., White, O., et al. 1995, The minimal gene complement of *Mycoplasma genitalium*, *Science*, **270**, 397–403.
6. Blattner, F.R., Plunkett, G. III, Bloch, C.A., et al. 1997, The complete genome sequence of *Escherichia coli* K-12, *Science*, **277**, 1453–74.
7. Kunst, F., Ogasawara, N., Moszer, I., et al. 1997, The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*, *Nature*, **390**, 249–56.
8. Weiner, J. III, Herrmann, R. and Browning, G.F. 2000, Transcription in *Mycoplasma pneumoniae*, *Nucleic Acids Res.*, **28**, 4488–96.
9. Friis, N.F. 1975, Some recommendations concerning primary isolation of *Mycoplasma suipneumoniae* and *Mycoplasma flocculare* a survey, *Nord. Vet. Med.*, **27**, 337–9.
10. Sambrook, J. and Russell, D.W. 2001, *Molecular Cloning: A Laboratory Manual*. Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY.
11. Bensing, B.A., Meyer, B.J. and Dunny, G.M. 1996, Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of *Enterococcus faecalis*, *Proc. Natl Acad. Sci. USA*, **93**, 7794–9.
12. Schneider, T.D. and Stephens, R.M. 1990, Sequence logos: a new way to display consensus sequences, *Nucleic Acids Res.*, **18**, 6097–100.
13. Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. 2004, WebLogo: a sequence logo generator, *Genome Res.*, **14**, 1188–90.
14. MacLellan, S.R., MacLean, A.M. and Finan, T.M. 2006, Promoter prediction in the rhizobia, *Microbiology*, **152**, 1751–63.
15. Hawley, D.K. and McClure, W.R. 1983, Compilation and analysis of *Escherichia coli* promoter DNA sequences, *Nucleic Acids Res.*, **11**, 2237–55.
16. Helmann, J.D. 1995, Compilation and analysis of *Bacillus subtilis* sigma A-dependent promoter sequences: evidence for extended contact between RNA polymerase and upstream promoter DNA, *Nucleic Acids Res.*, **23**, 2351–60.
17. Grech, B., Maetschke, S., Mathews, S. and Timms, P. 2007, Genome-wide analysis of Chlamydiae for promoters that phylogenetically footprint, *Res. Microbiol.*, **158**, 685–93.
18. Vogel, J., Axmann, I.M., Herzel, H. and Hess, W.R. 2003, Experimental and computational analysis of transcriptional start sites in the cyanobacterium *Prochlorococcus* MED4, *Nucleic Acids Res.*, **31**, 2890–9.
19. Wosten, M.M., Boeve, M., Koot, M.G., van Nuenen, A.C. and van der Zeijst, B.A. 1998, Identification of *Campylobacter jejuni* promoter sequences, *J. Bacteriol.*, **180**, 594–9.
20. Defrance, M., Janky, R., Sand, O. and van, H.J. 2008, Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences, *Nat. Protoc.*, **3**, 1589–603.
21. van Helden, J. 2003, Regulatory sequence analysis tools, *Nucleic Acids Res.*, **31**, 3593–6.
22. Thomas-Chollier, M., Sand, O., Turatsinze, J.V., et al. 2008, RSAT: regulatory sequence analysis tools, *Nucleic Acids Res.*, **36**, W119–27.
23. Hall, T.A. 1999, BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, *Nucleic Acids Symp. Ser.*, **41**, 95–8.
24. Bailey, T.L. and Elkan, C. 1994, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
25. Hertz, G.Z. and Stormo, G.D. 1999, Identifying DNA and protein patterns with statistically significant alignments of multiple sequences, *Bioinformatics*, **15**, 563–77.

26. Medina-Rivera, A., breu-Goodger, C., Thomas-Chollier, M., Salgado, H., Collado-Vides, J. and van, H.J. 2011, Theoretical and empirical quality assessment of transcription factor-binding motifs, *Nucleic Acids Res.*, **39**, 808–24.
27. Turatsinze, J.V., Thomas-Chollier, M., Defrance, M. and van, H.J. 2008, Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules, *Nat. Protoc.*, **3**, 1578–88.
28. Voskuil, M.I., Voepel, K. and Chambliss, G.H. 1995, The -16 region, a vital sequence for the utilization of a promoter in *Bacillus subtilis* and *Escherichia coli*, *Mol. Microbiol.*, **17**, 271–9.
29. Voskuil, M.I. and Chambliss, G.H. 2002, The TRTGn motif stabilizes the transcription initiation open complex, *J. Mol. Biol.*, **322**, 521–32.
30. Cases, I., Ussery, D.W. and de Lorenzo, V. 2003, The sigma54 regulon (sigmulon) of *Pseudomonas putida*, *Environ. Microbiol.*, **5**, 1281–93.
31. Turnbough, C.L. Jr. 2011, Regulation of gene expression by reiterative transcription, *Curr. Opin. Microbiol.*, **14**, 142–7.
32. Huerta, A.M., Francino, M.P., Morett, E. and Collado-Vides, J. 2006, Selection for unequal densities of sigma 70 promoter-like signals in different regions of large bacterial genomes, *PLoS Genet.*, **2**, e185.
33. Jarmer, H., Larsen, T.S., Krogh, A., Saxild, H.H., Brunak, S. and Knudsen, S. 2001, Sigma A recognition sites in the *Bacillus subtilis* genome, *Microbiology*, **147**, 2417–24.
34. Mitchell, J.E., Zheng, D., Busby, S.J. and Minchin, S.D. 2003, Identification and analysis of 'extended -10' promoters in *Escherichia coli*, *Nucleic Acids Res.*, **31**, 4689–95.
35. Guell, M., van Noort, V., Yus, E., et al. 2009, Transcriptome complexity in a genome-reduced bacterium, *Science*, **326**, 1268–71.
36. Petersen, L., Larsen, T.S., Ussery, D.W., On, S.L. and Krogh, A. 2003, RpoD promoters in *Campylobacter jejuni* exhibit a strong periodic signal instead of a -35 box, *J. Mol. Biol.*, **326**, 1361–72.
37. Hook-Barnard, I.G. and Hinton, D.M. 2007, Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters, *Gene Regul. Syst. Bio.*, **1**, 275–93.
38. Dekhtyar, M., Morin, A. and Sakanyan, V. 2008, Triad pattern algorithm for predicting strong promoter candidates in bacterial genomes, *BMC Bioinformatics*, **9**, 233.
39. Miroslavova, N.S. and Busby, S.J. 2006, Investigations of the modular structure of bacterial promoters, *Biochem. Soc. Symp.*, **73**, 1–10.
40. Orsini, G., Igonet, S., Pene, C., et al. 2004, Phage T4 early promoters are resistant to inhibition by the anti-sigma factor AsiA, *Mol. Microbiol.*, **52**, 1013–28.
41. Sinoquet, C., Demey, S. and Braun, F. 2008, Large-scale computational and statistical analyses of high transcription potentialities in 32 prokaryotic genomes, *Nucleic Acids Res.*, **36**, 3332–40.
42. Adams, C., Pitzer, J. and Minion, F.C. 2005, In vivo expression analysis of the P97 and P102 paralog families of *Mycoplasma hyopneumoniae*, *Infect. Immun.*, **73**, 7784–7.
43. Gardner, S.W. and Minion, F.C. 2010, Detection and quantification of intergenic transcription in *Mycoplasma hyopneumoniae*, *Microbiology*, **156**, 2305–15.
44. Siqueira, F.M., Schrank, A. and Schrank, I.S. 2011, *Mycoplasma hyopneumoniae* transcription unit organization: genome survey and prediction, *DNA Res.*, **18**, 413–22.
45. Reznikoff, W., Bertrand, K. and Donnelly, C. et al. 1987, *RNA Polymerase and the Regulation of Transcription*. In: Reznikoff, W., Burgess, R., Dahlberg, J., et al. (eds), Elsevier: New York, pp. 105–13.
46. Goodrich, J.A. and McClure, W.R. 1991, Competing promoters in prokaryotic transcription, *Trends Biochem. Sci.*, **16**, 394–7.
47. Brodolin, K., Zenkin, N., Mustaev, A., Mamaeva, D. and Heumann, H. 2004, The sigma 70 subunit of RNA polymerase induces lacUV5 promoter-proximal pausing of transcription, *Nat. Struct. Mol. Biol.*, **11**, 551–7.
48. Nickels, B.E., Mukhopadhyay, J., Garrity, S.J., Ebright, R.H. and Hochschild, A. 2004, The sigma 70 subunit of RNA polymerase mediates a promoter-proximal pause at the lac promoter, *Nat. Struct. Mol. Biol.*, **11**, 544–50.
49. Weiner, J. III, Zimmerman, C.U., Gohlmann, H.W. and Herrmann, R. 2003, Transcription profiles of the bacterium *Mycoplasma pneumoniae* grown at different temperatures, *Nucleic Acids Res.*, **31**, 6306–20.
50. Madsen, M.L., Nettleton, D., Thacker, E.L., Edwards, R. and Minion, F.C. 2006, Transcriptional profiling of *Mycoplasma hyopneumoniae* during heat shock using microarrays, *Infect. Immun.*, **74**, 160–6.
51. Madsen, M.L., Nettleton, D., Thacker, E.L. and Minion, F.C. 2006, Transcriptional profiling of *Mycoplasma hyopneumoniae* during iron depletion using microarrays, *Microbiology*, **152**, 937–44.
52. Schafer, E.R., Oneal, M.J., Madsen, M.L. and Minion, F.C. 2007, Global transcriptional analysis of *Mycoplasma hyopneumoniae* following exposure to hydrogen peroxide, *Microbiology*, **153**, 3785–90.
53. Oneal, M.J., Schafer, E.R., Madsen, M.L. and Minion, F.C. 2008, Global transcriptional analysis of *Mycoplasma hyopneumoniae* following exposure to norepinephrine, *Microbiology*, **154**, 2581–8.