UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
CENTRO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

IDENTIFICAÇÃO DE REGIÕES PROMOTORAS DE *Mycoplasma hyopneumoniae*

Tese de Doutorado

Shana de Souto Weber

Porto Alegre, junho de 2012

UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
CENTRO DE BIOTECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM BIOLOGIA CELULAR E MOLECULAR

IDENTIFICAÇÃO DE REGIÕES PROMOTORAS DE *Mycoplasma hyopneumoniae*

Tese submetida ao Programa de Pós-Graduação em Biologia Celular e Molecular do Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul como requisito parcial para a obtenção do grau de Doutor em Ciências.

Shana de Souto Weber

Orientadora: Profª Drª Irene Silveira Schrank

Porto Alegre, junho de 2012

Este trabalho foi desenvolvido no Laboratório de Organismos Diazotróficos, situado no Centro de Biotecnologia da Universidade Federal do Rio Grande do Sul. Financiamento: FAPERGS, CNPq e CAPES.

“Sem saber que era impossível, foi lá e fez.”

Jean Cocteau

Dedico esta tese à minha mãe, Marlene, e ao meu namorado, Fernando, que me incentivaram (com veemência) a ingressar nessa jornada; e ao meu pai, José Pedro, e à minha irmã, Gabriela, que sempre me apoiaram, independentemente das minhas escolhas.

AGRADECIMENTOS

À professora Irene Silveira Schrank, por ser uma orientadora tão paciente, dedicada e generosa; e por ter oportunizado a minha formação em seu laboratório.

Ao professor Augusto Schrank, por num belo dia, quando eu estava estressada com experimentos de *primer extension* (radioativo), ter perguntado: – Não dá para fazer isso com RACE?

Ao professor Sérgio Ceroni da Silva, por ter me dado as primeiras noções sobre bioinformática.

Aos professores Arnaldo Zaha, Henrique Ferreira e Marilene Vainstein, por terem me acompanhado e apoiado ao longo de todos estes anos.

À comissão de acompanhamento, composta pelos professores Augusto Schrank e Henrique Ferreira, pelas sugestões e críticas importantes.

À banca examinadora, formada pelos professores, Ana Tereza Ribeiro de Vasconcelos, Arnaldo Zaha e Edmundo Carlos Grisard, por terem aceitado avaliar este trabalho.

Aos meus colegas e ex-colegas de laboratório: Beatriz – minha companheirinha de bancada e de reciclagem de lixo; Clarissa – minha primeira colega de laboratório, Débora – amiga querida, parceira de limpeza e sempre minha companhia para ficar no laboratório até tarde; Dieime – minha ex-aluna de biomol, que faz o bolo de cenoura mais gostoso do mundo; Fernanda – estagiária meiguinha, que me "incomodou" durante um ano; Franciele – que, mesmo estando sempre de "saltinho", passou a usar tênis, virando minha parceira também no futsal; Luciano – por ter suportado bem a minha felicidade futebolística desde 2006 (...); Karyne e Maicon – tão animados que contagiavam qualquer um; Ricardo – o otimista, quem me ensinou as primeiras técnicas; e Scheila – aquela que agora habita a minha ex-bancada, agradeço pela amizade e cooperação.

Ao pessoal dos laboratórios 210, 217 e 220, em especial àqueles que convivem comigo há tanto tempo, Ângela, Broetto, Caru, Charley, Juli, Lívia, Roberta por serem sempre tão solícitos, por continuamente zelarem pelo bom funcionamento e organização dos laboratórios; ainda, agradeço à Bianca, pelos importantes cultivos de *M. hyopneumoniae*.

Aos funcionários do CBiot, principalmente, ao laboratorista, "Seu" Milton, por ter facilitado muito o nosso dia-a-dia; aos secretários do PPGBCM, Luciano e Sílvia, por serem tão prestativos, queridos e animados; ao "faz-tudo", Sr. Othelo, pelos concertos e pela alegria matinal.

Aos meus tão queridos companheiros de faculdade, pela amizade e carinho.

Aos meus familiares oficiais, os Weber de Cachoeira do Sul e os Souto de Florianópolis; e aos meus familiares emprestados, os Almeida/Darol e os Hayashi/Sant'Anna de Porto alegre, que sabem o trabalho que esta tese me deu, e torceram para que tudo desse certo.

Aos meus pais, Zé Pedro e Marlene, alegrias da minha vida, por me proporcionarem o conforto de um lar feliz, por serem tão especiais e amorosos.

À minha irmã querida, Gabi, que é um exemplo de determinação para mim.

Ao meu Ferdi, por ser o melhor companheiro, amigo, colega, que eu jamais imaginei ter; por sempre estar ao meu lado, mesmo quando não sou a melhor companhia. Por ter sido essencial para que esta tese chegasse ao fim... por ter me encorajado a encarar meus resultados, pelas tantas discussões que me ajudaram a ter certeza dos meus pontos de vista, por acreditar nesse trabalho e em mim.

À Universidade Federal do Rio Grande do Sul pelo ensino gratuito e de qualidade e ao apoio financeiro da FAPERGS, CAPES e CNPq.

ÍNDICE

ABREVIATURAS, SÍMBOLOS E UNIDADES

αI	subunidade alfa I da RNA polimerase
αII	subunidade alfa II da RNA polimerase
A	adenina
β	subunidade beta da RNA polimerase
β′	subunidade beta linha da RNA polimerase
b	base
°C	graus Celsius
C	citosina
cDNA	DNA complementar
CDS	seqüência de DNA codificante (DNA *coding sequence*)
CIP	fosfatase intestinal de bezerro (*calf intestinal phosphatase*)
CO_2	dióxido de carbono
dCDF	função de distribuição cumulativa decrescente (*decreasing cumulative distribution function*)
DNA	ácido desoxirribonucléico
dNTP	trifosfato de desoxirribonucleosídeos
FPR	taxa de falso-positivos (*false positive rate*)
g	gravidade
G	guanina
IPTG	isopropil-β-D-tiogalactopiranosídeo
IUPAC	International Union of Pure and Applied Chemistry
kb	quilobase
LOO	deixar um fora (*leave-one-out*)

µg	micrograma
min	minuto
µl	microlitro
ml	mililitro
mm	milímetro
mM	milimolar
NCBI	National Center for Biotechnology Information
NWD	diferença de valor de peso normalizada (*normalized weight score difference*)
ω	subunidade ômega da RNA polimerase
ORF	fase aberta de leitura (*open reading frame*)
pb	par de base
PCR	reação em cadeia da polimerase (*polymerase chain reaction*)
PES	pneumonia enzoótica suína
PFM	matriz de freqüência posicional (*position frequency matrix*)
pmol	picomol
PNK	cinase de polinucleotídeos (*polynucleotide kinase*)
PPM	matriz de probabilidade posicional (*position probability matrix*)
PSSM	matriz de pontuação posição-específica *(position-specific scoring matrix)*
PWM	matriz de peso posicional (*position weight matrix*)
RBS	sítio de ligação do ribossomo (*ribosomal binding site*)
RI	região intergênica
RLM-RACE	amplificação rápida das extremidades 5' dos cDNAs mediada por RNA ligase (*RNA ligase-mediated rapid amplification of 5´ cDNA ends*)
RNA	ácido ribonucléico
RNAP	RNA polimerase

ROC	características de operação do receptor *(receiver operating characteristic)*
rpm	revoluções por minuto
rRNA	RNA ribossômico
RSAT	Regulatory Sequence Analysis Tools
σ	subunidade sigma da RNA polimerase
SAP	fosfatase alcalina de camarão (*shrimp alkaline phosphatase*)
seg	segundo
T	timina
TAP	pirofosfatase ácida do tabaco (*tobacco acid pyrophosphatase*)
Tm	temperatura de fusão
tRNA	RNA transportador
TSS	sítio de início de transcrição (*transcriptional start site*)
U	unidade
UT	unidade de transcrição
UTR	região não traduzida (*untranslated region*)
vp	valor de peso
WD	diferença do valor de peso (*weight score difference*)

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Mycoplasma hyopneumoniae é uma das menores bactérias encontradas na natureza, apresentando genoma altamente reduzido e ausência de parede celular. Este organismo é o agente causador da pneumonia enzoótica suína, a qual apresenta distribuição mundial, causando importantes perdas econômicas. Na última década, várias espécies de *Mycoplasma* tiveram seus genomas completamente seqüenciados, incluindo quatro cepas de *M. hyopneumoniae*. Apesar da grande quantidade de dados gerados, pouco se sabe sobre as seqüências nucleotídicas que controlam a expressão gênica nestes microrganismos. A grande variabilidade encontrada nas regiões promotoras, o baixo conteúdo de GC presente no genoma e a carência de promotores experimentalmente caracterizados, são fatores que dificultam o reconhecimento *in silico* das seqüências reguladoras no gênero *Mycoplasma*. Assim sendo, este trabalho tem como objetivo identificar seqüências nucleotídicas envolvidas com o início da transcrição em *M. hyopneumoniae*, gerando dados que possibilitem a construção de uma matriz capaz de fazer a predição de promotores nesta espécie. Inicialmente, os sítios de início de transcrição (TSSs) de 23 genes de *M. hyopneumoniae* foram definidos. Os resultados mostraram que os TSSs identificados localizavam-se entre 2 e 144 pb de distância do início dos genes, sendo compostos em sua grande maioria por um resíduo de adenosina. Um padrão semelhante ao elemento −10 de promotores σ^{70} foi encontrado a montante dos TSSs. No entanto, não foi possível identificar conservação de uma provável região −35, porém, um sinal periódico AT-rico foi observado. Aproximadamente metade dos genes analisados continha o motivo 5'-TRTG-3', que é idêntico ao elemento −16, comumente encontrado em bactérias gram-positivas. A partir da determinação dos promotores, foi construída uma matriz de pontuação posição-específica que foi utilizada para localizar promotores putativos a montante de todas as seqüências codificantes (CDSs) de *M. hyopneumoniae*. Duzentos e um sinais foram encontrados associados a 169 CDSs. A maioria destas seqüências estava localizada até 100 nucleotídeos de distância dos códons de iniciação. Este estudo mostra que o número de seqüências promotoras preditas no genoma de *M. hyopneumoniae* é mais freqüente que o esperado ao acaso, indicando que a maioria das seqüências detectadas são provavelmente sítios de ligação funcionais.

ABSTRACT

Mycoplasma hyopneumoniae is one of the smallest bacteria found in nature, presenting a small genome and absence of cell wall. It is present in the majority of swine herds throughout the world, and is considered an important pathogen in the swine industry. In the last decade, many species of this genus had their genome completely sequenced, including four strains of *M. hyopneumoniae*. Nevertheless, very little is understood of the nucleotide sequences that control transcription initiation in these microorganisms. Like its relatives, *M. hyopneumoniae* lacks several major regulators of gene expression, including two component regulatory systems and multiple σ factors, thus it appears that the signals for promotion and regulation of transcription may differ significantly from other bacteria. In addition, the low GC content and the dearth of experimentally characterized promoters in this genus severely limit the recognition of the controlling sequences. Therefore, this study aims to identify nucleotide sequences involved in transcription initiation in *M. hyopneumoniae*, and thus generate data to enable the construction of a matrix capable of predicting promoters in this species. Initially, the transcription start sites (TSSs) of 23 genes of *M. hyopneumoniae* were experimentally defined. The results showed that the identified TSSs were located between 2 and 144 bp away from the gene starts, being composed mostly of an adenosine residue. 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'-TRTG-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.

1. INTRODUÇÃO

1.1 Transcrição bacteriana

A transcrição é o processo em que a informação contida na seqüência de DNA de um gene é sintetizada em RNA, dando origem a todas as moléculas de RNA necessárias ao funcionamento celular, tais como os RNAs mensageiros, estruturais e regulatórios. Nessa primeira etapa da expressão gênica, é determinada a maquinaria molecular necessária para diferenciação celular, morfogênese e adaptação de qualquer organismo (GHOSH *et al*., 2010), sendo, portanto, o nível mais efetivo em que a expressão de um gene pode ser regulada (VON HIPPEL, 1998).

O processo de síntese de RNA bacteriano, bastante estudado em *Escherichia coli*, consiste, basicamente, em 3 estágios: iniciação, alongamento e terminação.

Durante a iniciação, a RNA polimerase holoenzima (associada à σ^{70}) liga-se especificamente aos dois hexanucleotídeos conservados do promotor, nas posições −35 e −10 relativas ao sítio de início de transcrição, para formar o complexo fechado. A subseqüente separação das fitas do DNA, nas proximidades da região −10, resulta no complexo aberto e no início da transcrição. Depois de aproximadamente 12 nucleotídeos de RNA terem sido sintetizados, a holoenzima sofre uma mudança conformacional significativa, que leva, simultaneamente, à perda do contato com o promotor, liberação do fator σ e formação de um complexo de alongamento (BORUKHOV & NUDLER, 2003; WANG *et al*., 2011). A RNA polimerase (RNAP) prossegue o alongamento do transcrito até que sejam encontradas seqüências que indicam a terminação da transcrição. Essas seqüências formam estruturas secundárias no RNA nascente, que independentemente (terminação rho-independente) ou com o auxílio da ligação do fator Rho (terminação rho-dependente), fazem com que a RNAP

interrompa a adição de ribonucleotídeos, libere o transcrito completo e dissocie-se do DNA molde (WANG *et al*., 2011).

Conforme referido acima, a transcrição tem seu início e término determinados pelo reconhecimento de seqüências promotoras e terminadoras, respectivamente. A seqüência transcrita, delimitada por esses sítios, é definida como unidade de transcrição (UT). Portanto, cada UT origina uma única molécula de RNA, a qual pode ser composta por um (RNA monocistrônico) ou mais genes (RNA policistrônico) (BALLEZA *et al*., 2009). A co-transcrição de genes em uma mesma UT é comum nas bactérias, estando freqüentemente associada a genes cujos produtos estão envolvidos em processos biológicos intimamente relacionados (ERMOLAEVA *et al*., 2001). Dessa forma, os genes podem ser regulados em conjunto, sendo controlados coordenadamente através de um mesmo promotor.

A modulação da transcrição pode ocorrer nos seus diferentes estágios, porém, é durante a iniciação que a maioria dos eventos regulatórios acontece (GHOSH *et al*., 2010). Três elementos são fundamentais nesse estágio: a RNA polimerase, o fator de transcrição σ e as seqüências promotoras.

1.1.1 RNA polimerase

Presente em todas as etapas da transcrição, a RNA polimerase é considerada o componente central desse processo. Esta enzima consiste em um núcleo catalítico composto por cinco polipeptídios: αI e αII, β, β′, e ω, cujas seqüências, estruturas e funções são evolutivamente conservadas (DARST, 2001).

O sítio ativo da RNAP é formado a partir das grandes subunidades β e β′. A união destas é mantida por um dímero composto pelas duas subunidades α, sendo, portanto, responsável pela montagem do núcleo catalítico da enzima. A pequena subunidade ω, não tem

participação direta na transcrição, mas aparentemente funciona como uma chaperona que auxilia no dobramento de β′ (BROWNING & BUSBY, 2004).

Embora a RNAP, enquanto núcleo catalítico, seja competente para sintetizar RNA a partir de um DNA molde, ela é incapaz de iniciar a transcrição nos sítios apropriados (BORUKHOV & SEVERINOV, 2002). Para tanto, nas bactérias, é necessária a participação de um único polipeptídeo, o fator σ, que se liga ao núcleo resultando na formação da RNA polimerase holoenzima, a qual consegue reconhecer as seqüências promotoras e iniciar a transcrição (DARST, 2001).

1.1.2 Fator σ

O fator σ bacteriano possui três funções principais: garantir o reconhecimento de seqüências promotoras específicas, posicionar a RNA polimerase holoenzima no promotor alvo, e facilitar o relaxamento do DNA dupla-fita próximo ao sítio de início de transcrição (WOSTEN, 1998a).

Tipicamente, as bactérias possuem ao menos um fator σ primário, essencial para a viabilidade celular. Ele é responsável pela transcrição da maioria dos genes, particularmente daqueles expressos durante a fase de crescimento exponencial da célula (WOSTEN, 1998a; ÖSTERBERG *et al*., 2011). Os fatores σ primários de diferentes espécies são bastante similares, apresentando quatro regiões principais (σ1, σ2, σ3 e σ4), das quais são notavelmente mais conservadas as regiões envolvidas com reconhecimento dos elementos promotores (LONETTO *et al*., 1992; ÖSTERBERG *et al*., 2011). Apesar da semelhança funcional e estrutural, podem ser encontrados sob diferentes denominações: σ^{70} em *E. coli*, MysA em *Mycobacterium* spp., HdrB em *Streptomyces* spp., e SigA ou σ^{A} em *Bacillus subtilis* e outras bactérias gram-positivas (WOSTEN, 1998a).

Além do fator σ primário, a maioria das espécies contém diversos fatores σ não-

essenciais, os quais, quando combinados com a RNAP, formam holoenzimas que são capazes de reconhecer diferentes tipos de promotores, ativando conjuntos de genes específicos. Em *E. coli*, por exemplo, existem os fatores σ^{32}, que direcionam a síntese de genes, cujos produtos estabilizam ou renaturam proteínas; σ^{54}, que mobilizam genes envolvidos com a assimilação de componentes nitrogenados; σ^{S}, que aciona genes requeridos durante a fase estacionária ou em casos de estresse. Logo, a associação de fatores σ alternativos apropriados com o núcleo da RNAP permite às células ajustarem, rapidamente, o padrão de transcrição, de modo a otimizar o metabolismo celular em resposta às mudanças nas condições ambientais e nos sinais celulares (MOONEY *et al*., 2005; ÖSTERBERG *et al*., 2011).

O número genes que codificam fatores σ varia bastante de um organismo para outro: de apenas um nas espécies de *Mycoplasma* (*e.g*. FRASER *et al*., 1995), até mais de 60 em *Streptomyces coelicolor* (BENTLEY *et al*., 2002). Bactérias simbiontes e parasitas aparentemente requerem um número menor de fatores de transcrição próprios (MITTENHUBER, 2002), havendo, portanto, uma grande correlação entre o número de genes que codificam fatores σ e a diversidade de ambientes em que os organismos podem viver (BROWNING & BUSBY, 2004).

1.1.3 Promotores gênicos

Os promotores são seqüências de DNA responsáveis por determinar as regiões que devem ser transcritas em um genoma. Reconhecidos pela RNA polimerase holoenzima, eles indicam o sítio onde o início da transcrição deve ocorrer, além de influenciarem a afinidade de ligação da holoenzima, bem como a taxa de formação do complexo aberto (VAN HIJUM *et al*., 2009). Quatro seqüências diferentes podem ser identificadas em promotores reconhecidos pela RNAP associada a fatores σ primários: −10, −35, −10 estendido e elemento UP **[FIG. 1.1]** (BROWNING & BUSBY, 2004).

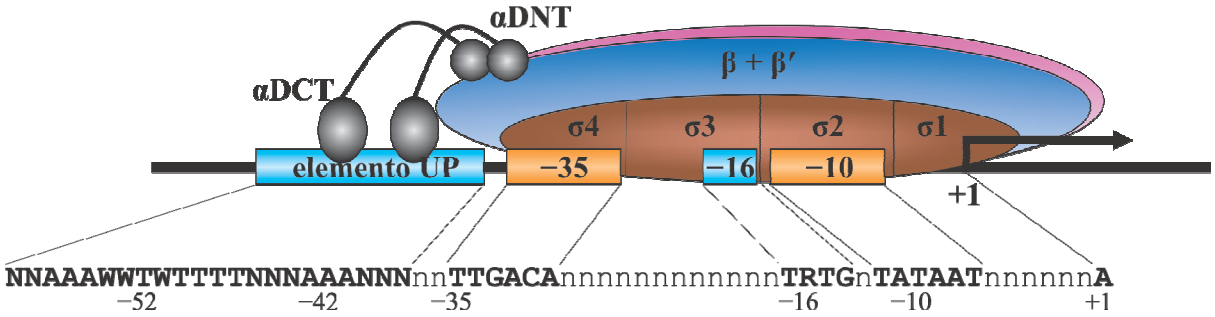


FIGURA 1.1 Elementos promotores e a interação com a RNA polimerase e o fator σ^{70}.
Ilustração mostrando as diferentes interações entre os elementos promotores e a RNA polimerase holoenzima. As fitas de DNA são representadas pela barra preta; nela estão destacados os elementos −10 e −35, em laranja, e os elementos −16 (−10 estendido) e UP, em azul-claro. A RNAP holoenzima é representada pelas subunidades β e β' coloridas em azul e rosa, respectivamente, pelos domínios carboxi-terminal (αDCT) e amino-terminal (αDNT) das subunidades α, identificados em cinza, e pelos diferentes domínios de σ, que estão em marrom. Abaixo, em detalhe, está a fita-senso mostrando as seqüências consenso dos elementos UP, −35, −16 e −10, além do DNA espaçador existente entre elas, representado pelas letras n (qualquer nucleotídeo). O sítio de início de transcrição, que freqüentemente é uma adenina, está indicado (+1). Adaptado de Browning & Busby (2004).

Dentre os elementos promotores a região −10 é, normalmente, a mais conservada (LISSER & MARGALIT, 1993), sendo necessária para o reconhecimento e ligação da RNAP holoenzima ao promotor, além de ser essencial para a formação do complexo aberto (HOOK-BARNARD & HINTON, 2007). O elemento −10 consiste em um hexanucleotídeo cuja seqüência consenso é $^{-12}$TATAAT^{-7} **[FIG. 1.1]**. Este elemento está centralizado 10 pb a montante do sítio de início de transcrição (+1), porém a distância entre eles pode variar de 4 a 8 pb (HAWLEY & MCCLURE, 1983; SHULTZABERGER *et al*., 2007). Reconhecida pelo domínio σ2 da subunidade σ da RNAP holoenzima **[FIG. 1.1]**, essa seqüência promotora pode ser reconhecida tanto como dupla-fita, indicando a localização do promotor, como simples-fita, estabilizando a enzima durante o processo de desnaturação do promotor (HOOK-BARNARD & HINTON, 2007).

O elemento −35 também interage com o fator σ, porém, diferentemente, é reconhecido pelo domínio σ4 **[FIG. 1.1]** e, somente como DNA dupla-fita (HOOK-BARNARD & HINTON, 2007). Assim como o elemento −10, −35 também é um hexanucleotídeo, tendo como consenso a seqüência $^{-35}$TTGACA^{-30} **[FIG. 1.1]**. Centralizado 35 pb a montante do +1, normalmente está localizado a 17±1 pb da região −10 (HAWLEY & MCCLURE, 1983).

Embora esse elemento seja importante para a identificação do promotor e para a ligação inicial da RNAP holoenzima, sua ocorrência não é imprescindível (HOOK-BARNARD & HINTON, 2007).

A terceira seqüência promotora reconhecida pelo fator σ^{70} é o elemento −10 estendido. Localizado a um par de base do hexanucleotídeo −10, tem como principal determinante o motivo $^{-15}$TG^{-14} (MITCHELL *et al*., 2003). Nas bactérias gram-positivas, a seqüência conservada costuma ser maior – $^{-17}$TRTG^{-14} (R= A ou G) –, sendo chamada de elemento −16 **[FIG. 1.1]** (VOSKUIL & CHAMBLISS, 1998). Em ambos os casos, essa região é especificamente contatada pelo domínio σ3 **[FIG. 1.1]** (BROWNING & BUSBY, 2004), fazendo com que a transcrição seja estimulada tanto pelo aumento da taxa de associação da RNAP holoenzima ao promotor, como pela estabilização do complexo aberto (VOSKUIL & CHAMBLISS, 2002). Essa interação é crítica para a transcrição de alguns promotores que possuem baixa conservação do consenso de um dos dois hexanucleotídeos (MITCHELL *et al*., 2003).

O reconhecimento e a atividade de promotores também podem ser afetados por seqüências AT-ricas localizadas imediatamente a montante da região −35, conhecidas como elementos UP **[FIG. 1.1]** (HOOK-BARNARD & HINTON, 2007). Com cerca de 20 pb de comprimento e, apesar de não serem tão conservados quanto os hexanucleotídeos, os elementos UP apresentam o consenso $^{-59}$NNAAAWWTWTTTTNNNAAANNN^{-38}, onde W = A ou T, e N = qualquer base. Essa seqüência ainda pode ser dividida em duas sub-regiões distintas, distal e proximal, as quais estão centralizadas nas posições −52 e −42 pb, respectivamente **[FIG. 1.1]**. Cada região dessas é reconhecida pelo domínio C-terminal de uma das duas subunidades α da RNAP holoenzima, estimulando a transcrição (ESTREM *et al*., 1999).

A combinação desses elementos, bem como a variação de suas seqüências,

determinam os diferentes níveis basais de transcrição, uma vez que afetam interações específicas do promotor com as subunidades σ e α da RNAP holoenzima. Embora elementos promotores com seqüências próximas aos consensos funcionem mais eficientemente, não existem promotores naturais nos quais todos esses sinais estejam presentes e sejam perfeitos. Isso porque, um promotor totalmente igual ao consenso ligaria a RNAP tão fortemente que impediria a transição da fase de iniciação da transcrição para a de alongamento (BROWNING & BUSBY, 2004). Também, promotores menos parecidos com a seqüência canônica proporcionariam a oportunidade de serem regulados. Sendo assim, os promotores teriam evoluído para conter um número ótimo de contatos para que ele não seja apenas reconhecido, mas para que também permita que a regulação e a transcrição aconteçam (HOOK-BARNARD & HINTON, 2007).

1.2 O gênero *Mycoplasma*

O gênero *Mycoplasma* é composto por alguns dos mais simples e menores organismos a se auto-replicarem. As bactérias desse táxon distinguem-se fenotipicamente de outros procariotos, principalmente, por não possuírem parede celular, característica que as faz pertencer à classe Mollicutes – do Latim: mollis, mole; cútis, pele (RAZIN *et al*., 1998). As primeiras espécies foram descritas há aproximadamente 70 anos, totalizando até o momento 123 espécies conhecidas (http://www.bacterio.cict.fr/m/mycoplasma.html).

Não existem micoplasmas de vida livre, estando amplamente distribuídos no reino animal como parasitas de mamíferos, aves, répteis, anfíbios e peixes (PITCHER & NICHOLAS, 2005). Normalmente, estão aderidos à superfície extracelular de células e tecidos do hospedeiro, embora já tenham sido descritas algumas espécies ocupando o interior de células eucarióticas (LO *et al*., 1993; BASEMAN *et al*., 1995). Podem ser patogênicos ou

apenas fazerem parte da microbiota natural do trato respiratório e urogenital (RAZIN, 2006). Nos homens, estão relacionados a doenças como asma, câncer, doenças auto-imunes, artrite e pneumonia (BASEMAN & TULLY, 1997).

No teste de Gram, coram como gram-negativas, uma vez que possuem apenas membrana plasmática, sem a proteção adicional de uma parede celular (RAZIN, 2006). Entretanto, filogeneticamente, os micoplasmas estão relacionados às bactérias gram-positivas, compartilhando, deste modo, um ancestral em comum com os gêneros *Streptococcus*, *Lactobacillus*, *Bacillus* e *Clostridium* (WOLF *et al*., 2004). Especificamente, a evolução teria ocorrido através de eventos de degeneração ou redução do genoma a partir de bactérias gram-positivas portadoras de parede celular e de genomas com baixo conteúdo de G+C (WOESE, 1987).

De acordo com os eventos evolutivos propostos, os genomas destes microrganismos são altamente reduzidos, variando entre 580 e 1350 kb (FRASER *et al*., 1995; SASAKI *et al*., 2002), apresentando baixo conteúdo de G+C, com variações entre 23 e 40% (WOESE, 1987). Além disso, as regiões intergênicas têm um maior conteúdo de A+T em relação às regiões codificantes – chegando a valores tão altos quanto 90% (DYBVIG & VOELKER, 1996). A distribuição de guanina e citosina, entre os genes, também irregular, ocorrendo em maior concentração nos genes que codificam rRNAs e tRNAs (FRASER *et al*., 1995). Como resultado desta composição atípica dos genomas, há o favorecimento da utilização de códons que contém adenina e timina (MUTO & OSAWA, 1987 *apud* BOVE, 1993) e, conseqüentemente, os micoplasmas possuem pouquíssimos códons GGN, CCN, GCN e CGN (RAZIN, 2006). Outra particularidade, que parece estar ligada a essa questão, é a utilização do códon UGA (OSAWA *et al*., 1992), que ao invés de ser utilizado como códon de terminação, conforme ao código genético universal, codifica para triptofano (YAMAO *et al*., 1985), característica que é igualmente encontrada no genoma das mitocôndrias.

Devido ao seu genoma reduzido, esses microrganismos não possuem várias das vias enzimáticas características da maioria das bactérias. Por exemplo, eles não apresentam vias *de novo* da biossíntese de purinas, um ciclo do ácido tricarboxílico completo, e um sistema de cadeia transportadora de elétrons mediada por citocromo (MANOLUKAS *et al*., 1988; POLLACK, 1992; FINCH & MITCHELL, 1992; FRASER *et al*., 1995). Isso estaria relacionado ao fato da maioria dos micoplasmas serem parasitas, geralmente, hospedeiro e tecido específicos (RAZIN *et al*., 1998).

1.3 *Mycoplasma hyopneumoniae*

M. hyopneumoniae é conhecido como o agente etiológico da pneumonia enzoótica suína (PES) – doença respiratória crônica, caracterizada por causar alta morbidade e baixa mortalidade (SOBESTIANSKY *et al*., 1999). A PES apresenta distribuição mundial, estando presente em quase todos os rebanhos suínos (MINION *et al*., 2004). Por determinar significante redução no ganho de peso, gastos com tratamento e, conseqüentemente, menores preços de venda das carcaças, a pneumonia enzoótica é causa de grandes perdas econômicas na produção intensiva de suínos (THACKER, 2006).

Esta bactéria é um patógeno extracelular que coloniza o trato respiratório através da aderência às células do epitélio ciliar (DEBEY & ROSS, 1994). Como demonstrado por Zielinski *et al*. (1990), essa aderência é essencial para que ocorra a colonização do organismo, sendo mediada principalmente por proteínas de membrana conhecidas como adesinas. O estabelecimento da infecção por *M. hyopneumoniae* resulta em ciliostase, perda dos cílios, morte das células epiteliais e inflamação aguda na traquéia, brônquios e bronquíolos, além de predispor o hospedeiro a infecções mais severas ocasionadas por patógenos secundários (CIPRIAN *et al*., 1988; DJORDJEVIC *et al*., 2004).

Devido o impacto econômico causado pela pneumonia enzoótica suína, há um grande empenho da comunidade científica em estudar *M. hyopneumoniae*. Prova disso é que quatro cepas (232, J, 7448 e 168) já tiveram seus genomas seqüenciados, fazendo desta, a espécie de micoplasma mais vezes seqüenciada (MINION *et al*., 2004; VASCONCELOS *et al*., 2005; LIU *et al*., 2011). A disponibilidade destes genomas tem propiciado o melhor entendimento da biologia molecular de *M. hyopneumoniae*, auxiliando em vários estudos referentes ao seu metabolismo e a sua patogenicidade.

1.4 Transcrição em *Mycoplasma* spp.

Além de apresentarem relevância médica ou veterinária, os micoplasmas também têm despertado interesse por possuírem genoma extremamente reduzido, fazendo com que sejam objetos de estudo bastante convenientes na determinação do conjunto mínimo de genes necessários para o estabelecimento de vida independente (GIL *et al*., 2004). Sendo assim, a partir da década de 1990, 23 espécies diferentes desse gênero tiveram seus genomas seqüenciados (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1).

O número de genes anotados, envolvidos com a transcrição, varia entre as seqüências genômicas analisadas. Na cepa 232 de *M. hyopneumoniae* são 11 genes e em *Mycoplasma penetrans* chegam a ser 23, correspondendo, respectivamente, a 1,6% e 2,2% do total de seqüências de DNA codificante (CDSs) (SASAKI *et al*., 2002; MINION *et al*., 2004). Entretanto, em *B. subtilis* – bactéria gram-positiva bastante estudada –, são encontrados 276 genes relacionados a este processo, equivalendo a 6,7% das seqüências codificadoras (SASAKI *et al*., 2002).

A RNA polimerase dos micoplasmas é semelhante a das eubactérias, sendo codificada pelos genes conservados *rpoA* (subunidade α), *rpoB* (subunidade β) e *rpoC*

(subunidade β'). Porém, somente um fator σ foi identificado (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1), diferentemente de *E. coli*, que tem no mínimo seis fatores σ (BLATTNER *et al*., 1997), e de *B. subtilis*, que contém no mínimo 18 deles (KUNST *et al*., 1997).

Embora não possuam fatores σ alternativos, trabalhos recentes, usando microarranjos, claramente mostram que *M. hyopneumoniae* pode regular seus genes em resposta às variações ambientais (MADSEN *et al*., 2006a; MADSEN *et al*., 2006b; SCHAFER *et al*., 2007; ONEAL *et al*., 2008). O mesmo foi observado em *Mycoplasma pneumoniae*, cujo transcriptoma revelou a ocorrência de ativação e repressão de genes, quando essa bactéria fora submetida a diferentes condições de cultivo (GÜELL *et al*., 2009).

A organização transcricional dos micoplasmas parece ser complexa. Através de estudo realizado com *M. hyopneumoniae*, Siqueira *et al*. (2011) sugeriram que, independentemente do tamanho das regiões intergênicas, genes adjacentes localizados na mesma fita de DNA – ou seja, com a mesma orientação –, seriam transcritos em uma única molécula de RNA. Desta forma, a transcrição iniciaria a partir de um promotor a montante do primeiro gene, e prosseguiria até que houvesse um gene na fita oposta (orientação oposta). Como conseqüência, *M. hyopneumoniae* apresentaria longos RNAs policistrônicos, muitas vezes compostos por genes com funções aparentemente não relacionadas.

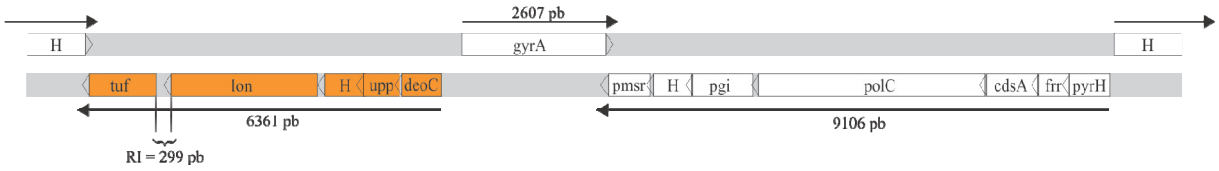


FIGURA 1.2 Organização transcricional dos genes de *Mycoplasma hyopneumoniae*.
Segmento do genoma de *M. hyopneumoniae* 7448 mostrando a organização transcricional proposta por Siqueira *et al*. (2011). Genes adjacentes, com a mesma orientação, seriam transcritos em uma única molécula de RNA, independentemente do tamanho das regiões intergênicas. Note que a unidade transcricional experimentalmente definida apresenta uma região intergênica (RI) relativamente longa entre os genes *tuf* e *lon*, e que ela é composta por genes aparentemente não relacionados. As unidades de transcrição (UT) estão identificadas por setas pretas, juntamente com seu tamanho. Os genes estão identificados por seus nomes ou pela letra H (gene hipotético). Em laranja: genes que compõem a UT experimentalmente definida. Em branco: genes que compõem UTs preditas.

A análise do transcriptoma de *M. pneumoniae*, além de evidenciar ser freqüente a heterogeneidade de genes compondo um mesmo transcrito, também mostrou que as unidades transcricionais (operons) podem ser subdivididas em unidades de transcrição menores (suboperons), implicando em uma alta taxa de transcritos alternativos. Interessantemente, foi observado que genes que eram separados nesses transcritos menores tendiam a pertencer a diferentes categorias funcionais. Essa geração de UTs alternativas seria resultado da modulação da transcrição tanto dos genes internos, quanto dos genes localizados no início ou no final dos transcritos (GÜELL *et al*., 2009); e indicaria a ocorrência de sítios de início e de terminação da transcrição internos.

Os mecanismos envolvidos na terminação da transcrição dos micoplasmas ainda não foram bem estabelecidos. Enquanto alguns autores sugerem que a terminação rho-independente é o principal mecanismo de terminação da transcrição (HOON *et al*., 2005; GÜELL *et al*., 2009), resultados de outros estudos indicam que os grampos de terminação não participariam deste processo (WASHIO *et al*., 1998; GARDNER & MINION, 2010). Recentemente, Gardner & Minion (2010) identificaram alguns potenciais grampos de terminação em *M. hyopneumoniae*, porém, verificaram que estes possuíam conteúdo de G+C inferior ao adequado para viabilizar uma terminação rho-independente típica (FARNHAM & PLATT, 1981; GARDNER & MINION, 2010). Nesse mesmo trabalho, foi demonstrado que, em *M. hyopneumoniae*, mais de 90% das regiões intergênicas maiores que 50 pb eram transcritas, provavelmente em decorrência de um controle de terminação da transcrição não-estringente. A terminação abrupta da transcrição, que indicaria a presença de terminadores, parece não ocorrer com grande freqüência. Sendo assim, foi proposto que o término da transcrição ocorreria preferencialmente de forma gradual, ou seja, a RNA polimerase continuaria a transcrição além do fim do gene, e se desligaria da fita-molde gradativamente (GARDNER & MINION, 2010).

Embora a complexidade observada no processo de transcrição dos micoplasmas possa ser explicada, em parte, pela ocorrência de RNAs anti-senso (LLUCH-SENAR *et al*., 2007; GÜELL *et al*., 2009), certamente não pode ser atribuída à presença dos poucos fatores de transcrição preditos (*e.g*. oito em *M*. *pneumoniae*) (GÜELL *et al*., 2009; YUS *et al*., 2009). Durante a evolução redutiva, à qual estes organismos foram submetidos, vários mecanismos regulatórios foram perdidos (HALBEDEL *et al*., 2007). Conseqüentemente, não possuem diversos genes codificando ortólogos bacterianos convencionais, tais como: os múltiplos fatores σ, sistemas de dois componentes (*two-component system*) e o fator de terminação de transcrição Rho (FRASER *et al*., 1995). Portanto, aparentemente, os sinais que promovem e regulam a transcrição, nos micoplasmas, devem diferir significativamente de outras bactérias (WEINER *et al*., 2000).

1.4.1 Promotores gênicos

As seqüências nucleotídicas que controlam o início e a regulação da transcrição nos micoplasmas são pouco entendidas, uma vez que, até agora, apenas alguns promotores foram identificados e analisados (WALDO *et al*., 1999; WEINER *et al*., 2000; MUSATOVOVA *et al*., 2003; HALBEDEL *et al*., 2007; ZHANG & BASEMAN, 2011). Esta carência de informação muito prejudica a interpretação dos dados obtidos com o seqüenciamento dos genomas. A anotação de regiões promotoras nestes microrganismos é especialmente dificultosa, pois, como observado em *M*. *hyopneumoniae*, o conteúdo de A+T nas seqüências codificantes (~70%) é tão alto quanto nas regiões intergênicas (~80%) (VASCONCELOS *et al*., 2005).

Weiner *et al*. (2000), com o intuito de aumentar a quantidade de informações disponíveis sobre os promotores de micoplasmas, determinaram os sítios de início de transcrição de 22 genes de *M*. *pneumoniae*. Estes sítios identificados e outros 10 sítios de

início de transcrição previamente descritos foram alinhados e, através da análise das 50 bases localizadas imediatamente à montante destes foi possível encontrar um forte consenso na região −10, enquanto que, na região −35 foi obtido apenas um consenso fraco. Várias possíveis regiões −10 foram identificadas: TA(AGT)AAT, TAA(GT)AT, TACTAT e TATTAA; e, considerando o conteúdo de G+C presente no genoma de *M. pneumoniae*, consensos similares aos das regiões −35 de *E. coli* foram detectados entre 15 e 20 bases de distância a 5' da região −10, apresentando uma seqüência TTGA relativamente conservada.

Em 2007, a região promotora do gene *ldh* de *M. pneumoniae* foi estudada *in vivo* (HALBEDEL *et al*., 2007). Fragmentos do gene contendo as prováveis regiões −10 e −35, de acordo com os consensos sugeridos anteriormente (WEINER *et al*., 2000), foram capazes de transcrever o gene *lacZ* no próprio *M. pneumoniae*, confirmando a sua atividade promotora e comprovando as predições feitas *in silico* por Weiner *et al*. (2000). A análise deste promotor, feita por meio da inserção de mutações pontuais, indica que, neste organismo, a região −10 é muito importante para transcrição, enquanto a região −35 proposta não apresenta a mesma relevância.

Nestes mesmos trabalhos, foi demonstrado que uma alta proporção de transcritos tem extremidades 5' heterogêneas, ocorrendo, em menor freqüência, entre as bases 1 e 4 a 5' do sítio de início de transcrição principal (WEINER *et al*., 2000). Ainda, grande parte dos RNAs mensageiros mostrou não possuir a região 5'-UTR que poderia conter o sítio RBS, resultado diferente do encontrado em outras bactérias (WEINER *et al*., 2000; HALBEDEL *et al*., 2007). A seqüência RBS (que é uma seqüência complementar à extremidade 3' do rRNA 16S) apenas pode ser encontrada em aproximadamente 80 genes de *M. pneumoniae* (WEINER *et al*., 2000).

Diferentes mecanismos de regulação envolvendo alterações reversíveis nas seqüências promotoras têm sido descritos em algumas espécies de *Mycoplasma*, sendo

responsáveis por controlar a expressão de várias famílias gênicas de lipoproteínas. Dentre estes eventos, podemos citar: expansão e contração de segmentos contíguos de resíduos de adeninas entre as regiões −10 e −35 (YOGEV *et al*., 1991), expansão e contração de repetições de trinucleotídeos a 5' da região promotora (GLEW *et al*., 1998), e inversões de DNA sítio-específicas envolvendo o promotor e a região RBS do gene (BHUGRA *et al*., 1995; NOORMOHAMMADI *et al*., 2000). Portanto, a identificação de promotores é fundamental para que o controle da expressão gênica nesses microrganismos seja bem entendido.

1.5 Ferramentas utilizadas no estudo de promotores de *Mycoplasma* spp.

1.5.1 *In vivo*

As tentativas iniciais para caracterizar funcionalmente as regiões promotoras de micoplasmas *in vivo* foram realizadas em *E. coli* (KNUDTSON & MINION, 1994; DHANDAYUTHAPANI *et al*., 1998). Porém, provavelmente devido ao DNA genômico rico em A+T dos micoplasmas, foi constatado que seqüências não reconhecidas como promotores nestas bactérias apresentavam falsa atividade promotora em *E. coli*, evidenciando, assim, a importância de se investigar a regulação gênica no próprio organismo de origem (KNUDTSON & MINION, 1994).

Halbedel & Stulke (2006), com o propósito de analisar as seqüências envolvidas no início da transcrição de *M. pneumoniae*, construíram o vetor pGP353. Este plasmídeo permite a clonagem de fragmentos contendo prováveis seqüências promotoras em frente à CDS do gene *lacZ*, o qual codifica a enzima β-galactosidade, uma das mais populares enzimas-repórter. Esse sistema-repórter foi escolhido por gerar resultados rápidos, tanto

qualitativos, através da visualização da atividade enzimática em colônias, quanto quantitativos, em ensaios usando como substrato cromatogênico o o-nitrofenil-D-galactopiranosídeo (MILLER, 1972 *apud* HALBEDEL & STULKE, 2006). Além disso, sistemas-repórter baseados no gene *lacZ* já foram estabelecidos em outros molicutes, tais como *Acholeplasma oculi*, *Mycoplasma pulmonis*, *Mycoplasma arthritidis* e *Mycoplasma capricolum* (KNUDTSON & MINION, 1994; DYBVIG *et al*., 2000; JANIS *et al*., 2005).

A análise molecular dos mecanismos envolvidos na regulação da transcrição em *M*. *hyopneumoniae* ainda não é possível devido à carência de sistemas-repórter adequados que possam ser utilizados para o estudo de suas regiões promotoras *in vivo*. Visando suprir essa deficiência é que, em nosso laboratório, foi desenvolvido o plasmídeo pOSTM, o qual possui origem de replicação de *M*. *hyopneumoniae*, e tem o gene de resistência à tetraciclina como gene-repórter (LOPES, 2007). Além disso, a utilização de pGP353 também foi avaliada com este objetivo (REOLON, 2007). Através da resistência aos antibióticos foi possível observar o isolamento de transformantes para ambos plasmídeos, no entanto, nenhum deles se manteve estável nas células de *M*. *hyopneumoniae*.

1.5.2 *In vitro*

A transcrição *in vitro* foi uma das primeiras metodologias utilizadas no estudo de promotores de *Mycoplasma* spp. (GAFNY *et al*., 1988; HYMAN *et al*., 1988). Nestes dois trabalhos, onde foram pesquisadas as regiões 5' dos operons de rRNA de *M*. *capricolum* e *M*. *pneumoniae*, a transcrição *in vitro* foi realizada com a RNA polimerase de *E*. *coli*. Entretanto, como mencionado anteriormente, os resultados obtidos por Knudtson & Minion (1994) sugerem que a RNA polimerase de *E*. *coli* pode reconhecer seqüências que não possuem atividade promotora nos micoplasmas.

De acordo com esse resultado, é evidente a necessidade de utilização da RNA polimerase holoenzima de *Mycoplasma* spp. Outras metodologias, como *gel mobility assay* e *footprinting*, também requereriam o emprego da enzima espécie-específica para definir as verdadeiras regiões envolvidas com o início da transcrição nestes organismos. No entanto, ainda não estão disponíveis protocolos para a purificação da holoenzima nativa e nem para sua expressão heteróloga, o que implicaria na expressão de diferentes subunidades e na utilização de uma estratégia de substituição dos múltiplos códons TGA.

1.5.3 *In silico*

Vários programas desenvolvidos para fazer o reconhecimento de promotores estão disponíveis. No entanto, estes têm sido predominantemente baseados na grande quantidade de informações disponíveis para *E. coli* (WEINER *et al*., 2000), não considerando o contexto biológico das seqüências analisadas (KIM & SIM, 2005). Este fato dificulta a identificação de promotores em micoplasmas, tendo em vista que possuem um conteúdo de G+C menor do que o apresentado pela bactéria *E. coli*. Nestes casos foi sugerida a adaptação dos algoritmos através de ajustes que levem em consideração o conteúdo de guanina/citosina do genoma em questão, visando, assim, possibilitar a utilização destes em diferentes espécies bacterianas (HERTZ & STORMO, 1996).

Poucos trabalhos têm se dedicado a produzir ferramentas de bioinformática destinadas especificamente a encontrar seqüências promotoras em micoplasmas. Além de uma matriz de peso gerada por Weiner *et al*. (2000) a partir de matrizes baseadas em dados sobre promotores de *E. coli* (HERTZ & STORMO, 1996), há alguns anos, foi desenvolvido um sistema utilizando inteligência artificial (VALIATI, 2006). Porém, neste foi verificada uma baixa capacidade preditiva, não identificando, com a certeza desejada, novos possíveis promotores. Portanto, embora existam ferramentas de bioinformática que auxiliam na

determinação de seqüências regulatórias, estas são limitadas pela falta de promotores experimentalmente caracterizados nesse gênero (WEINER *et al*., 2000; VALIATI, 2006).

1.6 Análise *in silico* de promotores bacterianos

A anotação dos genes e de suas funções em um genoma fornece informações relacionadas à organização gênica, ao metabolismo, às funções celulares, proporcionando, no entanto, pouco conhecimento sobre as vias que regulam a expressão gênica (MÜNCH *et al*., 2011). Nesse contexto, a predição de promotores se torna importante, uma vez que possibilita inferir quais elementos regulatórios controlam a expressão de cada gene (HUERTA & COLLADO-VIDES, 2003). Além disso, a indicação de potenciais promotores pode proporcionar uma melhor anotação do genoma (WANG & BENHAM, 2006), além de servir como ponto de partida para estudos experimentais (RANGANNAN & BANSAL, 2009; BLAND *et al*., 2010).

Apesar da predição de regiões codificantes ter progredido, os métodos *in silico* de identificação de promotores ainda não apresentam a mesma habilidade (QIU, 2003). Uma das principais dificuldades inerentes à predição desses elementos regulatórios é que suas seqüências são curtas e não são totalmente conservadas. Portanto, há uma grande probabilidade de que seqüências semelhantes aos promotores possam ser encontradas em regiões do genoma que não estejam relacionadas com a regulação do início da transcrição (KANHERE & BANSAL, 2005). Outro fator complicador é a diversidade existente entre os genomas bacterianos, o que torna difícil a detecção de sítios regulatórios conservados por homologia de seqüência (WANG & BENHAM, 2006).

1.6.1 Estratégias de predição

Basicamente, existem duas abordagens gerais para o reconhecimento de padrões de seqüências (*e.g.* seqüências promotoras). Uma delas chama-se descoberta de padrões (*pattern discovery*), sendo capaz de realizar a predição de motivos *ab initio*, ou seja, sem ter conhecimento prévio a respeito do sítio de ligação e da proteína reguladora correspondente. Esta metodologia baseia-se na idéia de que genes funcionalmente relacionados ou co-regulados compartilham as mesmas seqüências regulatórias, as quais ocorrem, portanto, com maior freqüência do que esperado ao acaso. Normalmente, os conjuntos de seqüências analisados costumam ser provenientes de genes co-expressos verificados em experimentos de microarranjos de DNA e *ChIP-on-chip*, ou de genes ortólogos de organismos relacionados (MÜNCH *et al*., 2011). Os motivos compartilhados por essas seqüências são identificados através de algoritmos, como MEME e ClustalW, que verificam se a freqüência de ocorrência de um motivo é ou não estatisticamente significativa (BAILEY & ELKAN, 1994; HERTZ & STORMO, 1999).

A outra abordagem é conhecida como correspondência de padrões (*pattern matching*). Essa faz uso de conhecimento prévio na forma de um padrão pré-determinado que pode ser atribuído a um regulador específico. O padrão é comumente estabelecido com base num perfil de sítios de ligação de fatores de transcrição definidos, para os quais existem dados experimentais disponíveis **[FIG. 1.1 A]**. Usando este conjunto de seqüências, um modelo probabilístico que descreve a degeneração do padrão é construído (MÜNCH *et al*., 2011). As ferramentas computacionais que empregam essa abordagem podem ser divididas em duas grandes categorias, baseada em seqüência (*string-based*) e baseada em matriz (*matrix-based*), dependendo de como o motivo está sendo representado (TURATSINZE *et al*., 2008).

[A] 40 sítios de ligação do fator de transcrição Anr

1	T	T	G	A	T	T	C	C	G	G	T	C	A	A
2	T	T	G	A	C	T	T	T	C	A	T	C	A	A
3	T	T	G	A	T	T	G	C	C	A	T	C	A	A
4	T	T	G	A	C	C	G	G	A	A	T	C	A	A
⋮								⋮						
40	T	T	G	A	T	G	T	C	G	A	T	C	A	A

[B] Consenso estrito

T	T	G	A	T	C	T	G	G	A	T	C	A	A

[C] Consenso degenerado

Y	T	G	H	Y	N	B	N	B	V	K	C	A	R

[D] Matriz de freqüência

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	2	0	0	29	0	6	0	10	6	20	0	0	40	35
C	5	0	0	7	17	17	15	10	15	3	1	40	0	0
G	0	0	40	0	3	10	9	18	16	17	2	0	0	5
T	33	40	0	4	20	7	16	2	3	0	37	0	0	0

[E] Matriz de probabilidade

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0,05	0,01	0,01	0,71	0,01	0,15	0,01	0,25	0,15	0,49	0,01	0,01	0,98	0,86
C	0,13	0,01	0,01	0,18	0,42	0,42	0,37	0,25	0,37	0,08	0,03	0,98	0,01	0,01
G	0,01	0,01	0,98	0,01	0,08	0,25	0,23	0,45	0,40	0,42	0,05	0,01	0,01	0,13
T	0,81	0,98	0,01	0,10	0,49	0,18	0,40	0,05	0,08	0,01	0,91	0,01	0,01	0,01

[F] Matriz de peso posicional

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	-1,52	3,71	-3,71	1,05	-3,71	-0,49	-3,71	0,00	-0,49	0,68	-3,71	-3,71	1,37	1,24
C	0,67	-3,71	-3,71	-0,35	0,52	0,52	0,40	0,00	0,40	-1,15	-2,10	1,37	-3,71	-3,71
G	-3,71	-3,71	1,37	-3,71	-1,15	0,00	-0,10	0,58	0,46	0,52	-1,52	-3,71	-3,71	-0,67
T	1,18	1,37	-3,71	0,88	0,68	-0,35	0,46	-1,52	-1,15	-3,71	1,29	-3,71	-3,71	-3,71

[G] Logo de seqüência

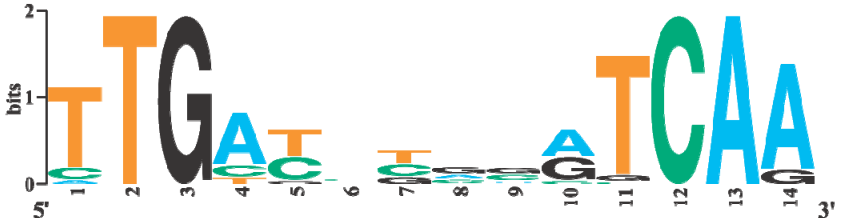


FIGURA 1.3 Diferentes representações do padrão apresentado por um conjunto de sítios de ligação. Representações dos sítios de ligação específicos para o fator de transcrição Anr de *Pseudomonas aeruginosa* (TRUNK *et al*., 2010). [A] Resumo de uma coleção de 40 sítios de ligação experimentalmente caracterizados do fator de transcrição Anr de *P. aeruginosa*. [B] Consenso estrito, indicando a base mais freqüente em cada posição. [C] Consenso degenerado, representando a existência de posições ambíguas, como por exemplo, a letra R, que indica que em determinada posição há maior probabilidade de serem encontradas A ou G. [D] Matriz de freqüência (PFM) (sem a adição dos *pseudo-counts*) mostrando a freqüência absoluta de cada nucleotídeo em cada posição. [E] Matriz de probabilidade (PPM) mostrando a freqüência relativa de cada nucleotídeo em determinada posição. Note que não existem valores iguais a zero, pois essa matriz foi calculada após a PFM ter sido corrigida com *pseudo-counts*. [F] Matriz de peso posicional (PWM ou PSSM) mostrando o logaritmo da divisão de cada probabilidade da PPM pela probabilidade de determinado nucleotídeo ser encontrado no *background*. [G] *Logo* de seqüência. Cada coluna representa uma posição do motivo, e as letras indicam que nucleotídeos são encontrados em determinada posição. A altura geral da pilha de nucleotídeos indica a conservação da seqüência naquela posição, e a altura de cada nucleotídeo em cada pilha indica a sua freqüência relativa nessa posição. Adaptado de Münch *et al*. (2011).

1.6.1.1 Correspondência de padrões

1.6.1.1.1 Predição baseada em seqüência

A predição com base em seqüência é a metodologia mais simples de descrever um motivo de ligação, uma vez que uma única seqüência resume a informação contida em uma coleção de sítios de ligação (TURATSINZE *et al*., 2008; MÜNCH *et al*., 2011). Esses sítios são alinhados e o nucleotídeo mais comum de cada posição específica é designado como consenso para representar a composição de nucleotídeos de cada coluna (WASSERMAN & SANDELIN, 2004). O consenso pode utilizar um alfabeto restrito ao código de quatro letras (A, C, G e T; consenso estrito) **[FIG. 1.1 B]**, ou utilizar o código de 15 letras da IUPAC, que permite representar posições ambíguas (consenso degenerado) **[FIG. 1.1 C]** (TURATSINZE *et al*., 2008). Embora a seqüência consenso seja um modo conciso de representar um conjunto de seqüências, e sirva para comparações visuais rápidas, ela não reflete de forma adequada a importância de cada base em cada posição/coluna (MÜNCH *et al*., 2011).

1.6.1.1.2 Predição baseada em matriz

Uma descrição mais precisa dos sítios de ligação é obtida através de modelos probabilísticos, os quais levam em conta a freqüência de cada nucleotídeo em cada posição de um motivo, não considerando somente as bases mais comuns de cada posição. Tais modelos são geralmente chamados de matriz de peso posicional (*position weight matrix*, PWM) ou matriz de pontuação posição-específica (*position-specific scoring matrix*, PSSM) (MÜNCH *et al*., 2011).

A PWM é comumente criada a partir de uma matriz de freqüência posicional (*position frequency matrix*, PFM) via uma matriz de probabilidade posicional (*position*

probability matrix, PPM). Com base no alinhamento de um conjunto de sítios de ligação, o número total de observações de cada nucleotídeo para cada posição é registrado, produzindo a PFM **[FIG. 1.1 D]**. A partir dessa matriz é gerada a PPM, que exibe dados de freqüência relativa, informando a probabilidade de cada nucleotídeo ser observado em determinada posição **[FIG. 1.1 E]** (WASSERMAN & SANDELIN, 2004). Essa probabilidade é, para alguns nucleotídeos, próxima ou igual a zero, o que é provavelmente resultado do número finito de sítios de ligação conhecidos que são utilizados para construir a matriz. Para evitar esse viés, devido a um pequeno tamanho amostral, um certo valor numérico – *pseudo-count*, é usualmente alocado para cada posição, e sua fração é adicioanada a cada elemento na PFM, para então calcular a PPM (NISHIDA *et al*., 2009).

Uma vez construída a PPM, cada probabilidade contida nessa matriz é dividida pela probabilidade do nucleotídeo em questão ser observado no *background* (por exemplo, o *background* pode ser o genoma ou as regiões intergênicas do organismo cujas seqüências serão analisadas pela matriz). O resultado dessa divisão é convertido à escala logarítmica, dando origem a PWM **[FIG. 1.1 F]**. Portanto, os valores de uma PWM são o logaritmo da razão entre duas probabilidades (TURATSINZE *et al*., 2008). A informação contida nas matrizes é normalmente visualizada em *logos* de seqüência, os quais fornecem uma representação gráfica intuitiva da importância de cada resíduo em cada posição do motivo **[FIG. 1.1 G]**. (SCHNEIDER & STEPHENS, 1990).

Com o auxílio de programas que façam a correspondência de padrões, a PWM é, então, utilizada na busca por potenciais sítios de ligação, atribuindo um valor quantitativo às seqüências analisadas (WASSERMAN & SANDELIN, 2004). Desta forma, o programa percorre a seqüência investigada através de uma "janela" de tamanho igual ao motivo designado pela matriz, deslocando-se uma base por vez. A cada segmento compreendido pela janela é atribuída uma pontuação específica. Essa pontuação corresponde à soma dos valores,

conferidos, pela PWM, a cada nucleotídeo em cada posição do segmento de seqüência, de acordo com a similaridade ao motivo representado por ela (TURATSINZE *et al*., 2008). Portanto, quanto maior for a pontuação, maior será a probabilidade de que determinado segmento de seqüência seja um sítio de ligação.

Conforme descrito acima, tanto seqüências funcionais como não-funcionais recebem uma pontuação. Seqüências regulatórias são comumente curtas (normalmente 6−18 pb), o número de sítios comprovados experimentalmente é muitas vezes limitado e, em muitos casos, o nível observado de conservação da seqüência é baixo. Conseqüentemente, a freqüência dos diferentes padrões previstos pela matriz no genoma é, muitas vezes, exageradamente alta (MÜNCH *et al*., 2011). Sendo assim, há a necessidade de definir um valor de corte que auxilie a discriminação entre sítios de ligação de fatores de transcrição biologicamente funcionais e sítios não-funcionais.

Embora subjetiva, a definição do valor de corte pode ser otimizada ao valer-se de parâmetros estatísticos. Ao menos dois parâmetros devem ser considerados nesse processo: a sensibilidade e a taxa de falso-positivos associadas a uma dada pontuação (MÜNCH *et al*., 2011). A sensibilidade (ou taxa de verdadeiro-positivos) é a proporção de verdadeiro-positivos identificados entre os sítios funcionais (falso-negativos e verdadeiro-positivos), enquanto, a taxa de falso-positivos informa a proporção de falsos positivos entres os sítios não-funcionais (falso-positivos e verdadeiro-negativos). Quanto menor a pontuação limiar definida, maior o número de predições falsas (falso-positivos). No entanto, se o limiar for muito alto, aumenta-se a chance de perder sítios biologicamente funcionais, ou seja, há um aumento do número de falso-negativos (MÜNCH *et al*., 2011). Portanto, é crucial definir um valor de corte que garanta uma taxa de falso-positivos razoavelmente baixa e que mantenha uma taxa de predições corretas (sensibilidade) satisfatória.

2. OBJETIVOS

2.1 Objetivos gerais

A estrutura dos promotores de *Mycoplasma* spp. ainda é pouco conhecida, limitando o entendimento da regulação gênica a partir dos dados obtidos com o seqüenciamento de seus genomas. Em *Mycoplasma hyopneumoniae*, esta falta de informação deve-se, em parte, a inexistência de ferramentas genéticas e computacionais. Assim sendo, este trabalho tem como objetivos a caracterização de promotores de *M. hyopneumoniae* e a construção de uma matriz de peso posicional espécie-específica que permita a predição dos promotores dessa bactéria.

2.2 Objetivos específicos

- Determinação experimental de sítios de início de transcrição de genes de interesse;
- Identificação de promotores através de uma abordagem *ab initio*;
- Caracterização dos promotores identificados;
- Construção e otimização de uma matriz de peso;
- Predição dos promotores de *M. hyopneumoniae*;
- Avaliação do desempenho da matriz na predição de promotores nas demais espécies de *Mycoplasma*.

3. MATERIAIS & MÉTODOS

3.1 Linhagens bacterianas e condições de crescimento

Mycoplasma hyopneumoniae 7448, procedente da Embrapa Suínos e Aves (Concórdia, SC), foi isolada de um suíno infectado em Lindóia do Sul, SC, Brasil. O cultivo foi feito em tubos Falcon de 15 ml contendo 5 ml de meio Friis (FRIIS, 1975), a 37°C por 48 horas, sob leve agitação.

Escherichia coli XL1-Blue (Invitrogen) foi mantida em meio sólido Luria-Bertani contendo 20 μg/ml de tetraciclina (SAMBROOK & RUSSELL, 2001). Quando transformada com plasmídeo pUC18 recombinante, foi cultivada em meio Luria-Bertani (sólido e líquido) contendo 100 μg/ml de ampicilina. Ainda, ao meio sólido foram adicionados 40 μg/ml de X-Gal (5-bromo-4-cloro-3-indolil-β-D-galactopiranosídeo) e 0,3 mM de IPTG (isopropil-β-D-tiogalactopiranosídeo). A cultura foi feita a 37°C por 16 horas, com agitação de 150 rpm (quando em meio líquido).

3.2 Seleção de genes para determinação dos sítios de início de transcrição

Os genes, que tiveram seus sítios de início de transcrição (*transcriptional start site*, TSS) determinados, foram selecionados com base em dois critérios: (I) ter função determinada, não podendo estar anotado como hipotético e (II) ter o gene a montante em orientação divergente **[FIG. 3.1]**.

A anotação do genoma de *M. hyopneumoniae* 7448, disponível no National Center of Biotechnology Information (NCBI) sob o código de acesso NC_007332, foi utilizada para análise do contexto gênico.

FIGURA 3.1 Contexto gênico que deve ser apresentado pelo gene selecionado para determinação do TSS. Região do genoma de *M. hyopneumoniae* 7448 contendo os genes *uvrC* e *dnaK*, os quais apresentam orientação divergente, sendo possíveis alvos de estudo.

3.3 Extração de RNA

O RNA total foi isolado a partir de 25 ml de cultura de *M. hyopneumoniae* 7448. As células foram coletadas por centrifugação a 3360 × *g* por 15 minutos e ressuspendidas em 1 ml de TRIzol (Invitrogen). A suspensão de células foi, então, processada de acordo com o protocolo fornecido pelo fabricante. Subseqüentemente, 50 µg de RNA foram tratadas com a DNase livre de RNase RQ1 (Promega), seguido de purificação e concentração com o *kit* NucleoSpin® RNA Clean-up XS (Macherey-Nagel).

3.4 Oligonucleotídeos e condições de termociclagem

O adaptador 5' RACE, os *primers* 5' RACE Outer e 5' RACE Inner foram fornecidos no *kit* First Choice RLM-RACE (Ambion). Os *primers* gene-específicos **[ANEXO 8.1]**, empregados no experimento de 5' RLM-RACE, foram projetados a partir das seqüências nucleotídicas anotadas no genoma da cepa 7448 de *M. hyopneumoniae* (número de acesso no NCBI: NC_007332). Para cada gene foram sintetizados dois *primers* gene-específicos, os

quais, juntamente com os *primers* 5' RACE Outer e Inner, formaram os pares de *primers* externos e internos necessários para a execução das *nested*-PCRs.

As *nested*-PCRs foram feitas no termociclador Mastercycler (Eppendorf) e utilizaram a técnica de *touchdown* **[TAB. 3.1]**.

TABELA 3.1 Condições de termociclagem

	Ciclos	Temperatura*	Tempo
Desnaturação inicial	1 ×	94°C	2 min
Amplificação	5 ×	94°C	30 seg
		(Tm + 6°C)	30 seg
		72°C	1 min
	5 ×	94°C	30 seg
		(Tm + 4°C)	30 seg
		72°C	1 min
	20 ×	94°C	30 seg
		Tm	30 seg
		72°C	1 min
Extensão final	1 ×	72°C	10 min

* Tm corresponde à temperatura de fusão estimada para o primer gene-específico.

Os *primers* M13 *forward* e *reverse* (Invitrogen) foram utilizados na confirmação dos clones, realizada através de PCR de colônia, e nas reações de seqüenciamento.

3.5 Purificação dos fragmentos, clonagem e transformação

A purificação dos produtos obtidos com os experimentos de 5' RLM-RACE foi feita a partir de gel de agarose, utilizando o *kit* NucleoSpin® Extract II (Macherey-Nagel). Depois de purificados, os fragmentos de DNA foram tratados com Klenow DNA polimerase (Fermentas) e com cinase de polinucleotídeos (*polynucleotide kinase*, PNK, Fermentas), e

foram clonados no vetor pUC18 clivado com SmaI e defosforilado com fosfatase alcalina de camarão (*shrimp alkaline phosphatase*, SAP, USB), seguindo as orientações dos fabricantes.

A ligação dos fragmentos ao vetor pUC18-SmaI, a transformação por eletroporação na cepa de *E. coli* XL1-Blue, a seleção dos clones, a confirmação destes por PCR de colônia e a extração de DNA plasmidial foram feitas usando métodos padrões descritos por Sambrook & Russell (2001).

3.6 Seqüenciamento automático e análise dos dados

Os clones foram seqüenciados utilizando o *kit* DYEnamic Et Dye Terminator Cycle Sequencing (Amersham Biosciences), desenvolvido para o seqüenciador automático MegaBACE 1000 DNA Analysis System (Healthcare), conforme o manual do fabricante.

Os resultados dos seqüenciamentos foram processados no pacote de programas Staden (STADEN *et al*., 2000) em conjunto com o programa PHRED (EWING *et al*., 1998). Após, a análise dos produtos seqüenciados e a identificação dos TSSs foram feitas através do alinhamento manual das seqüências obtidas nos seqüenciamentos com as seqüências gênica e intergênica localizadas na região 5' do gene correspondente.

3.7 Identificação dos sítios de início de transcrição

Para determinar os TSSs, foi empregada a técnica de amplificação rápida das extremidades 5' dos cDNAs mediada por RNA ligase (*RNA ligase-mediated rapid amplification of 5´ cDNA ends*, 5' RLM-RACE), a qual é baseada na estratégia descrita por Bensing *et al*. (1996) **[FIG. 3.2]**. Esta metodologia foi executada utilizando o *kit* First Choice RLM-RACE (Ambion), de acordo com o protocolo do fabricante, com exceção do tratamento

com a fosfatase intestinal de bezerro (*calf intestinal phosphatase,* CIP), o qual não foi realizado.

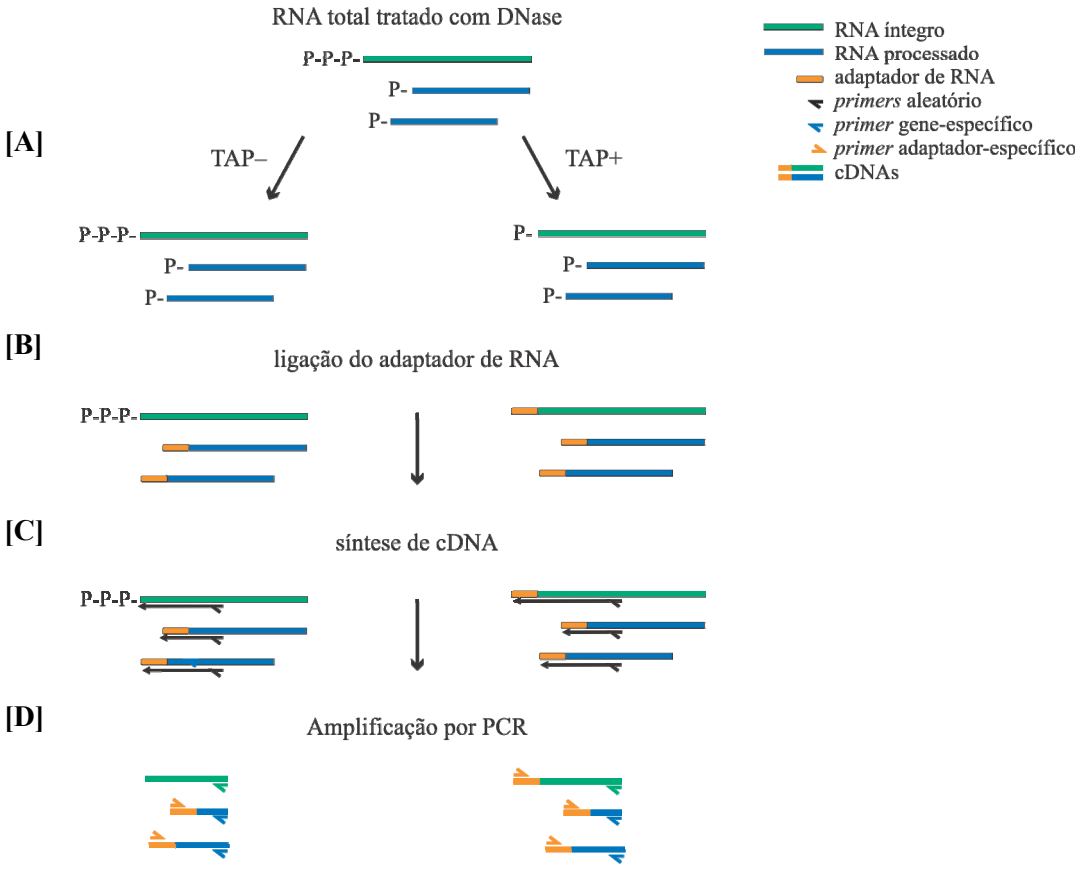


FIGURA 3.2 Metodologia utilizada na identificação dos TSSs – 5' RLM-RACE.
Esquema mostrando as diferentes etapas do processo. Três transcritos são mostrados neste exemplo, dois relacionados (um é o produto processado do outro) e um não relacionado. [A] As amostras duplicadas são tratadas ou não com a enzima TAP para converter 5' trifosfatos em monofosfatos. [B] Um oligonucleotídeo de RNA, de seqüência conhecida, é ligado às extremidades 5' monofosfatadas. [C] O cDNA é feito usando *primers* aleatórios. [D] Os cDNAs são amplificados com um *primer* de DNA correspondente a seqüência do adaptador de RNA ligado, e com um *primer* gene-específico. Os produtos de PCR são separados por eletroforese em gel de agarose. Os produtos diferenciais presentes na amostra tratada com TAP (fragmentos derivados de RNA não processado, ou seja, RNA 5' trifosfatado), mas ausentes na amostra não tratada, possuem o sítio de início de transcrição correto.

Resumidamente, foi feita uma reação de 16 µl contendo 10 µg RNA livre de DNA, tampão da pirofosfatase ácida do tabaco (*tobacco acid pyrophosphatase,* TAP) e 20 U de inibidor de RNase (Fermentas). Essa mistura foi dividida em duas alíquotas, das quais uma recebeu 2 µl da enzima TAP (reação TAP+) e a outra recebeu 2 µl de água (reação TAP−). Após a etapa de tratamento com TAP, as duas amostras (TAP+ e TAP−) foram processadas

de forma idêntica durante as reações de ligação do adaptador 5' RACE e de transcrição reversa.

Uma vez obtido o cDNA, três *nested*-PCRs foram feitas para cada gene: TAP+, TAP− e controle negativo. Todas as reações foram feitas em um volume total de 25 µl contendo 1,25 mM de $MgCl_2$, tampão da Taq 1×, 0,02 mM de cada dNTP, 1 U de Taq DNA polimerase (Ludwig Biotec), 10 pmol dos *primers* adaptador- e gene-específico e 0,5 µl de DNA molde. A PCR externa foi feita com cDNA como DNA molde, e utilizou os *primers* 5' RACE e gene-específico "outer". A PCR interna foi feita usando uma alíquota da PCR externa como DNA molde, e utilizou os *primers* 5' RACE e gene-específico "inner".

As amplificações foram realizadas usando a técnica PCR *touchdown*, e os produtos foram analisados em géis de agarose 1,2–2%. Os produtos presentes na amostra tratada com TAP (fragmentos derivados de RNA não-processado), mas ausentes na amostra não tratada, foram purificados e clonados **[SEÇÃO 3.5]**. Os clones foram analisados através da metodologia de PCR de colônia para confirmar a presença do inserto, sendo, então, seqüenciados.

3.8 *Logos* de seqüências

O *logos* de seqüências foram gerados através do site WebLogo (http://weblogo.berkeley.edu/) (SCHNEIDER & STEPHENS, 1990; CROOKS *et al*., 2004). Seqüências promotoras relacionadas ao fator σ^{70}, experimentalmente definidas, foram utilizadas de acordo com os alinhamentos propostos pelos seus respectivos autores **[ANEXO 8.2]**. Os seguintes números de seqüências promotoras foram empregados para gerar os *logos*: 25 sítios de *Sinorhizobium meliloti* (MACLELLAN *et al.*, 2006), 59 sítios de *E. coli* (HAWLEY & MCCLURE, 1983), 142 sítios de *Bacillus subtilis* (HELMANN, 1995), 41

sítios de *Chlamydia trachomatis* (GRECH *et al.*,2007), 35 sítios de *Mycoplasma pneumoniae* (WEINER *et al.*, 2000), 25 sítios de *Prochlorococcus marinus* (VOGEL *et al.*, 2003), 21 sítios de *Campylobacter jejuni* (WOSTEN *et al.*, 1998) e 23 sítios de *M. hyopneumoniae*. O tamanho do genoma e o conteúdo de G+C foram obtidos a partir dos genomas depositados no NCBI (http://www.ncbi.nlm.nih.gov/).

3.9 Cálculo da distância evolutiva entre fatores σ^{70}

A diferença entre os fatores σ^{70} de espécies de diferentes gêneros foi mensurada através do cálculo da estimativa de distância evolutiva existente entre seqüências. Especificamente, foram analisadas as regiões dos domínios de σ^{70} responsáveis por reconhecerem os elementos promotores.

Utilizando o programa MUSCLE (EDGAR, 2004), foram alinhadas as seqüências das proteínas σ^{70} dos seguintes organismos: *S. meliloti* (NP_386420.1), *E. coli* (NP_289642.1), *B. subtilis* (ZP_03592291.1), *C. trachomatis* (NP_220132.1), *M. pneumoniae* (NP_110040.1), *P. marinus* (NP_892614.1), *C. jejuni* (YP_179074.1) e *M. hyopneumoniae* (YP_287460.1). As seqüências foram obtidas a partir do NCBI; os códigos de acesso estão informados entre parênteses.

As regiões 2.4, 3.0 e 4.2 envolvidas diretamente no reconhecimento dos elementos promotores −10, −16 e −35, respectivamente, foram identificadas de acordo com descrito por Gruber & Gross (2003). Cada uma dessas regiões foi isolada e analisada separadamente pelo programa MEGA5 (TAMURA *et al*., 2011), o qual foi empregado para calcular a distância evolutiva existente entre cada região nos diferentes organismos. Nesta análise foi utilizado o modelo baseado em matriz JTT (Jones-Taylor-Thornton) (JONES *et al*., 1992), e todas as posições contendo lacunas foram eliminadas. As distâncias foram

calculadas em relação ao fator σ^{70} de *E. coli*, por este ser o mais bem caracterizado dentre os demais.

3.10 Identificação de padrões de seqüências

A procura por padrões de seqüências foi realizada na região a montante dos TSSs determinados pela metodologia de 5' RLM-RACE. Três programas diferentes foram utilizados: Local-Word-Analysis (DEFRANCE *et al*., 2008) do Regulatory Sequence Analysis Tools (RSAT) (http://rsat.ulb.ac.be/) (VAN HELDEN, 2003; THOMAS-CHOLLIER *et al*., 2008), Multiple Em for Motif Elicitation (MEME) (BAILEY & ELKAN, 1994) e WConsensus (HERTZ & STORMO, 1999).

Com o Local-Word-Analysis foram analisadas as primeiras 50 bases a montante dos TSSs em busca de motivos compostos por seis ou quatro nucleotídeos. Motivos de seis nucleotídeos foram localizados aplicando uma janela com largura fixa de 10 nucleotídeos, enquanto que, para os motivos de quatro nucleotídeos, utilizou-se uma janela fixa de 5 nucleotídeos. Em ambos os casos, o modelo de *background* empregado considerava a composição nucleotídica apenas das seqüências intergênicas localizadas a montante de todos os genes de *M. hyopneumoniae* 7448.

Com os programas MEME e WConsensus foram analisadas as primeiras 25 bases a montante dos TSSs. Nestas duas análises não foram pré-definidos o tamanho do motivo e largura da janela. O mesmo modelo de *background* (identificado como Bernoulli model) foi empregado, sendo calculado com base no mesmo conjunto de seqüências descritas acima.

3.11 Construção das PSSMs

Os motivos identificados com o programa Local-Word-Analysis, localizados entre 4 a 8 bases a montante do TSS, foram alinhadas manualmente com auxílio do programa BioEdit 7.0 (HALL, 1999). Este alinhamento foi usado para construir uma matriz de peso composta por 12 colunas. Além desta, outras duas matrizes, compostas por 14 e 16 colunas, foram obtidas através dos programas MEME e Wconsensus, respectivamente. Por fim, as matrizes foram reconstruídas sem a presença de seqüências repetidas.

3.12 Seqüências-alvo

Todas as seqüências foram adquiridas a partir do genoma completo de cada espécie de *Mycoplasma* através da utilização da ferramenta Retrive Sequence do RSAT.

3.12.1 Conjunto de seqüências nº 1

As seqüências empregadas na avaliação do desempenho das matrizes (cálculo da distribuição empírica) **[SEÇÃO 3.13.1]**, e na predição dos promotores **[SEÇÃO 3.15]** foram extraídas a partir de todos os genes codificantes de *M. hyopneumoniae* 7448 (código de acesso no NCBI: NC_007332). As 657 seqüências obtidas consistiam em até 250 bases a montante (coletando apenas a região não-codificante) e 50 bases a jusante do códon de iniciação anotado para cada gene.

3.12.2 Conjunto de seqüências nº 2

As seqüências utilizadas para avaliar o desempenho da matriz de 12 colunas nas diferentes espécies de *Mycoplasma* **[SEÇÃO 3.13.2]**, especificamente no cálculo da distribuição empírica, foram obtidas a partir de todos os genes codificantes de cada organismo. As seqüências consistiam em até 250 bases a montante (coletando apenas a região não-codificante) e 50 bases a jusante do códon de iniciação anotado para cada gene.

3.12.3 Conjunto de seqüências nº 3

As seqüências usadas para definição do valor de corte **[SEÇÃO 3.14]** foram obtidas a partir das seqüências descritas no conjunto nº1 **[SEÇÃO 3.12.1]**. Foram criados dois grupos: (i) o de "orientação correta", que consistia nas mesmas seqüências do conjunto nº1 e (ii) o de "orientação incorreta", que consistia no complemento reverso das mesmas seqüências do grupo de orientação correta. De ambos os grupos foram removidas seqüências localizadas entre genes transcritos divergentemente.

3.13 Avaliação da performance das PSSMs

A habilidade das PSSMs em identificar sítios de ligação funcionais reconhecidos pelo fator σ^{70} foi avaliada usando o programa Matrix-Quality (MEDINA-RIVERA *et al*., 2011) do RSAT.

Para cada avaliação foram obtidas distribuições de valores de peso teórica e empírica, a partir das quais foi calculada a diferença de valores de peso normalizada (*normalized weight difference*, NWD).

3.13.1 PSSMs de 12, 14 e 16 colunas aplicadas à *M. hyopneumoniae*

O desempenho das três PSSMs em identificar sítios promotores funcionais em *M. hyopneumoniae* foi avaliado. Para cada matriz, cinco distribuições teóricas e empíricas foram calculadas, empregando diferentes modelos de *background*. As distribuições empíricas foram obtidas a partir do conjunto de seqüências nº 1 **[SEÇÃO 3.12.1]**. As distribuições teóricas foram obtidas a partir de uma seqüência randômica gerada de acordo com o modelo de *background*. Ambas as distribuições foram computadas aplicando os seguintes parâmetros: um *pseudo-count* para correção da matriz, *pseudo-frequencies* ajustadas em 0,01 e, como modelo de *background*, foram testadas as ordens de Markov de 0 a 4, utilizando como base a composição nucleotídica de seqüências não-codificantes localizadas a montante de todos os genes da cepa 7448 de *M. hyopneumoniae*.

A análise comparativa do desempenho de cada matriz, associada a cada modelo de *background*, foi realizada por meio do cálculo do NWD. A partir dessa comparação, foram identificados a matriz e modelo de *background* que apresentaram a melhor capacidade preditiva.

Para a matriz de melhor performance foi feito um controle adicional, no qual as distribuições empíricas e teóricas da PSSM de 12 colunas foram comparadas com a média de dez PSSMs de colunas permutadas. Estas foram derivadas a partir da PSSM de 12 colunas através da ferramenta Permute-Matrix do RSAT.

3.13.2 PSSM de 12 colunas aplicada às diferentes espécies de *Mycoplasma*

O desempenho da PSSM de 12 colunas também foi avaliado na identificação de sítios promotores funcionais nas demais espécies do gênero *Mycoplasma*. A matriz foi aplicada em 21 outras espécies e em outras duas cepas de *M. hyopneumoniae*: *Mycoplasma*

agalactiae PG2, *Mycoplasma arthritidis* 158L3-1, *Mycoplasma bovis* PG45, *Mycoplasma capricolum* subsp. *capricolum* ATCC 27343, *Mycoplasma conjunctivae* HRC/581, *Mycoplasma crocodyli* MP145, *Mycoplasma fermentans* JER, *Mycoplasma gallisepticum* str. R(low), *Mycoplasma genitalium* G37, *Mycoplasma haemofelis* str. Langford 1, *Mycoplasma hominis* ATCC 23114, *Mycoplasma hyopneumoniae* 232, *Mycoplasma hyopneumoniae* J, *Mycoplasma hyorhinis* HUB-1, *Mycoplasma leachii* PG50, *Mycoplasma mobile* 163K, *Mycoplasma mycoides* subsp. *mycoides* SC str. PG1, *Mycoplasma penetrans* HF-2, *Mycoplasma pneumoniae* M129, *Mycoplasma pulmonis* UAB CTIP, *Mycoplasma putrefaciens* KS1, *Mycoplasma suis* KI3806, *Mycoplasma synoviae* 53.

A distribuição empírica de cada espécie foi obtida a partir da análise do conjunto de seqüências nº 2 **[SEÇÃO 3.12.2]**, enquanto a distribuição teórica foi obtida a partir de uma seqüência randômica gerada de acordo com o modelo de *background*. Ambas as distribuições foram computadas empregando os seguintes parâmetros: um *pseudo-count* para correção da matriz; *pseudo-frequencies* ajustadas em 0,01; e a ordem 1 de Makov como modelo de *background*, sendo calculada com base na composição nucleotídica de seqüências não-codificantes localizadas a montante de todos os genes do organismo em questão.

Para uma melhor visualização do desempenho da matriz em cada espécie, foram gerados gráficos de NWD.

3.14 Determinação do valor de corte

O valor de corte foi determinado através da abordagem descrita por Cases *et al*. (2003), a qual compara a distribuição dos valores de peso entre os promotores preditos a partir de seqüências de "orientação correta" **[SEÇÃO 3.12.3]**, com aqueles encontrados nas seqüências de "orientação incorreta" **[SEÇÃO 3.12.3]**. O valor de peso que apresentou uma considerável

redução de promotores putativos na orientação incorreta foi selecionado como o valor de corte.

A estimativa da taxa de falso-positivos (*false positive rate*, FPR) e da sensibilidade associadas a este valor de corte foram observadas através de curvas características de operação do receptor (*receiver operating characteristic*, ROC), geradas pelo Matrix-Quality. Uma validação do tipo *Leave-One-Out* (LOO) do conjunto de sítios usados para construir a matriz foi incluída.

3.15 Predição de promotores

Promotores putativos de *M. hyopneumoniae* foram procurados no conjunto de seqüências nº 1 **[SEÇÃO 3.12.1]** com a PSSM de 12 colunas usando o programa Matrix-Scan (TURATSINZE *et al*., 2008). Os parâmetros *pseudo-counts*, *pseudo-frequencies* e o modelo de *background* foram ajustados conforme feito no programa Matrix-Quality **[SEÇÃO 3.13]**, exceto pelo fato de que foi utilizada a ordem 1 de Markov.

4. RESULTADOS

4.1 Genes selecionados para determinação do sítio de início de transcrição

Com o objetivo de aumentar a probabilidade de encontrarmos os TSSs imediatamente a montante dos genes escolhidos, a seleção destes foi feita com base na função e contexto gênicos, conforme a anotação do genoma de *Mycoplasma hyopneumoniae* **[SEÇÃO 3.2]**. Portanto, o gene selecionado deveria apresentar orientação divergente do gene localizado a montante, garantindo que este não estaria dentro de uma unidade de transcrição, e possuir uma função definida, ou seja, não poderia estar classificado como hipotético, uma vez que a transcrição destes genes não é conhecida.

Trinta e quatro genes, pertencentes a diferentes categorias funcionais, foram selecionados **[TAB. 4.1]**. Dentre eles, podemos destacar a presença de genes de interesse no estudo da patologia e fisiologia de *M. hyopneumoniae*, tais como: genes que codificam proteínas envolvidas com citoaderência, como por exemplo, as adesinas P97 e P146; o gene MHP7448_0513, que codifica o antígeno 46K (P46); o gene *sipS*, que codifica uma sinal peptidase com potencial antigênico para imunodiagnóstico e vacinação (MOITINHO-SILVA *et al*., 2012); o gene *deoC*, primeiro gene da unidade transcricional que comporta o gene que codifica o fator de alongamento Tu (EF-Tu), altamente antigênico (PINTO *et al*., 2007); e o gene MHP7448_0225, primeiro gene da unidade de transcrição que codifica genes envolvidos no metabolismo de mio-inositol, os quais foram observados apenas em *M. hyopneumoniae* (KLEIN, 2008).

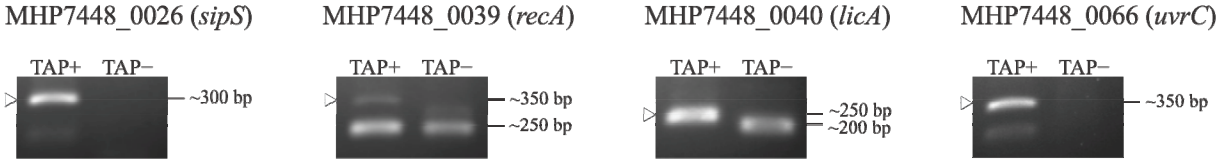
TABELA 4.1 Genes selecionados para identificação do sítio de início de transcrição

Locus	Gene	Posição Genômica	Descrição	Categorias Funcionais do COG
MHP7448_0026	*sipS*	32885−33325	signal peptidase I	Posttranslational modification, protein turnover, chaperones
MHP7448_0039	*recA*	52346−53335	recombinase A	DNA replication, recombination and repair
MHP7448_0040	*licA*	53566−54321	PTS system, lichenan-specific IIA component	Cell envelope biogenesis, outer membrane
MHP7448_0066	*uvrC*	85675−87192	excinuclease ABC subunit C	DNA replication, recombination and repair
MHP7448_0067	*dnaK*	87839−89641	molecular chaperone DnaK	Posttranslational modification, protein turnover, chaperones
MHP7448_0101	*clpB*	134736−136829	ATP-dependent protease binding protein	Posttranslational modification, protein turnover, chaperones
MHP7448_0161	*deoB*	199410−200606	phosphopentomutase	Carbohydrate transport and metabolism
MHP7448_0195	*rpsJ*	220229−220600	30S ribosomal protein S10	Translation, ribosomal structure and biogenesis
MHP7448_0198	0198	221969–225238	protein P97	-
MHP7448_0224	*glyA*	267185−268441	serine hydroxymethyltransferase	Amino acid transport and metabolism
MHP7448_0225	0225	268939−270408	methylmalonate-semialdehyde dehydrogenase	Energy production and conversion
MHP7448_0241	*secD*	287718−290315	bifunctional preprotein translocase subunit SecD/SecF	-
MHP7448_0272	0272	328638−331712	protein P97-like	-
MHP7448_0279	0279	338452−339270	transcriptional regulator	Transcription
MHP7448_0359	*glpK*	449002−450534	glycerol kinase	Energy production and conversion
MHP7448_0360	0360	450873−452138	P37-like ABC transporter substrate-binding lipoprotein	-
MHP7448_0427	*efp*	540785−541342	elongation factor P	Translation, ribosomal structure and biogenesis
MHP7448_0454	*acpD*	599048−599632	acyl carrier protein phosphodiesterase	Lipid metabolism
MHP7448_0490	*pgk*	640265−641479	phosphoglycerate kinase	Carbohydrate transport and metabolism
MHP7448_0505	0505	667597−670413	lipoprotein	-
MHP7448_0513	0513	680272−681531	46K surface antigen precursor	-
MHP7448_0521	*pepF*	689586−691430	oligoendopeptidase F	Amino acid transport and metabolism
MHP7448_0527	*deoC*	703355−704020	deoxyribose-phosphate aldolase	Nucleotide transport and metabolism
MHP7448_0528	*gyrA*	704416−707022	DNA gyrase subunit A	DNA replication, recombination and repair
MHP7448_0535	*pyrH*	715459−716169	uridylate kinase	Nucleotide transport and metabolism
MHP7448_0545	*ktrA*	726591−727286	potassium uptake protein	Inorganic ion transport and metabolism
MHP7448_0546	*ktrB*	727584−729068	potassium uptake protein	Inorganic ion transport and metabolism
MHP7448_0586	*nusA*	775502−777355	transcription elongation factor NusA	Transcription
MHP7448_0619	*rplJ*	825913−826410	50S ribosomal protein L10	Translation, ribosomal structure and biogenesis
MHP7448_0622	*dam*	831858−833507	DNA adenine methylase	DNA replication, recombination and repair
MHP7448_0647	*leuS*	866103−868475	leucyl-tRNA synthetase	Translation, ribosomal structure and biogenesis
MHP7448_0648	*uvrB*	868683−870662	excinuclease ABC subunit B	DNA replication, recombination and repair
MHP7448_0654	*prsA*	874822−875808	ribose-phosphate pyrophosphokinase	Amino acid transport and metabolism
MHP7448_0663	0663	890392−894372	adhesin like-protein P146	-

4.2 Amplificação e diferenciação da região 5' de transcritos primários e processados

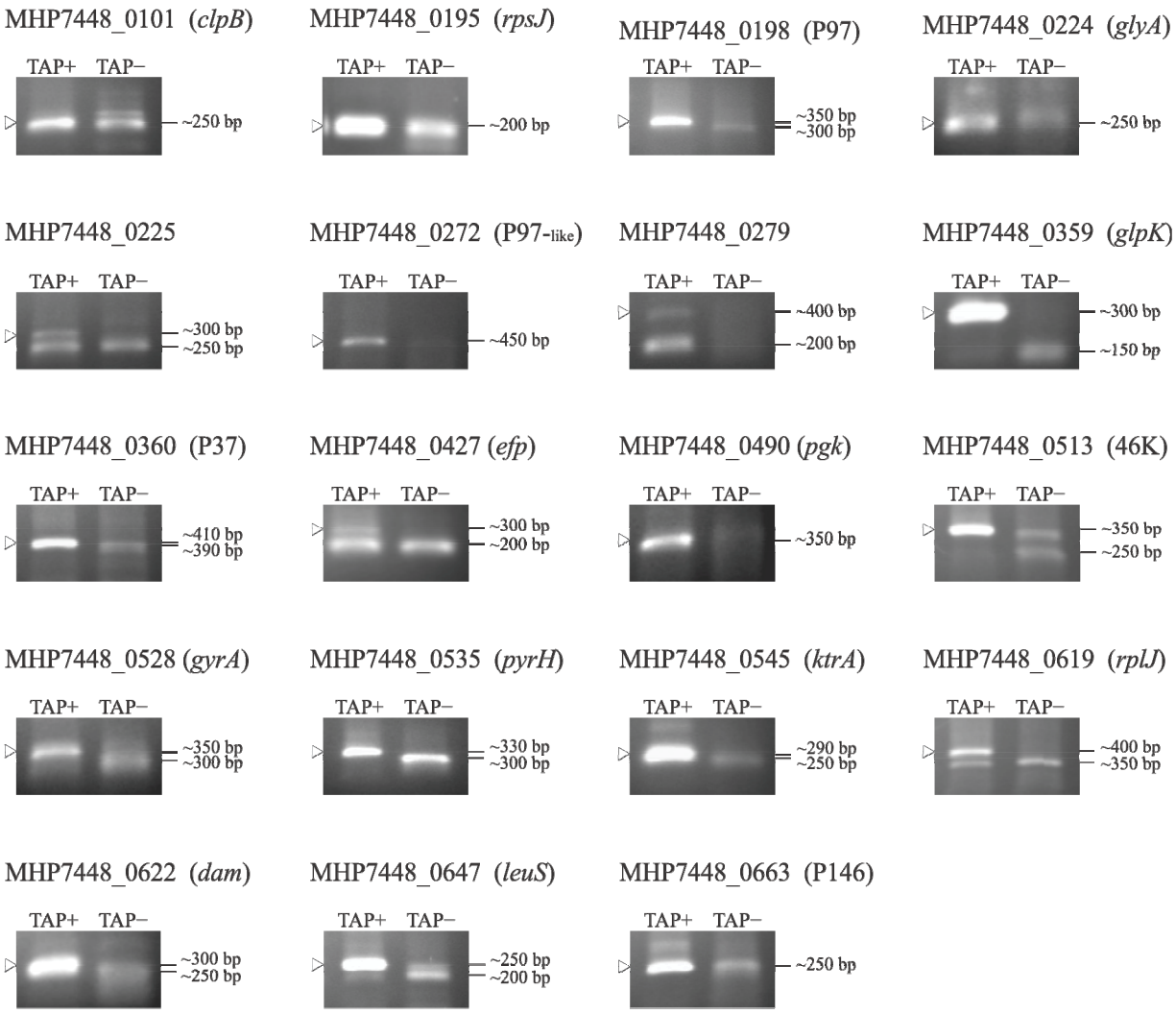
Para identificação dos TSSs corretos, foi empregada a metodologia de 5' RLM-RACE, a qual permite distinguir entre transcritos primários e processados de acordo com o estado de fosforilação de suas extremidades 5' – trifosfatadas ou monofosfatadas, respectivamente **[SEÇÃO 3.7 – FIG. 3.2]**. Essa diferenciação é feita através da comparação entre os produtos de 5' RLM-RACE oriundos de RNA tratado e de RNA não-tratado com a fosfatase TAP. Essa enzima transforma a extremidade 5' trifosfatada em monofosfatada, possibilitando que a partir das amostras tratadas com TAP sejam amplificados ambos transcritos, primários e processados, enquanto que, a partir das amostras que não foram tratadas, sejam amplificados apenas transcritos processados. Sendo assim, os produtos de amplificação provenientes de amostras de RNA tratadas com TAP devem apresentar, em análise por eletroforese em gel, um sinal específico, ou ao menos mais intenso, em comparação com os produtos originados de amostras de RNA que não foram tratadas (BENSING *et al*., 1996).

Dos 34 genes analisados, 23 apresentaram sinal específico ou mais intenso nos produtos de 5' RLM-RACE provenientes das amostras tratadas com TAP **[FIG. 4.1]**. Estes produtos foram purificados e clonados. Os demais genes, para os quais o resultado esperado não foi obtido, tiveram seus estudos descontinuados.



(parte 1 de 2)

FIGURA 4.1 Amplificação e diferenciação da extremidade 5' de transcritos primários e processados. Produtos de 5' RLM-RACE oriundos de RNA tratado (TAP+) e de RNA não-tratado (TAP−) com TAP. Os produtos da amostra tratada com TAP que foram purificados e clonados estão identificados por triângulos. Visualização em géis de agarose corados com brometo de etídio.



(parte 2 de 2)

FIGURA 4.1 Amplificação e diferenciação da extremidade 5' de transcritos primários e processados. Produtos de 5' RLM-RACE oriundos de RNA tratado (TAP+) e de RNA não-tratado (TAP−) com TAP. Os produtos da amostra tratada com TAP que foram purificados e clonados estão identificados por triângulos. Visualização em géis de agarose corados com brometo de etídio.

4.3 Determinação dos sítios de início de transcrição

A análise da região 5' dos transcritos de cada gene foi feita a partir do seqüenciamento de dez ou mais clones independentes, obtidos através do experimento de 5' RLM-RACE.

Os resultados dos seqüenciamentos revelaram que 12 genes possuíam um TSS bem definido. No entanto, em 11 genes, a extremidade 5' apresentou-se variável. Essa

variação, existente entre os transcritos de um mesmo gene, devia-se a alguns poucos nucleotídeos de diferença, os quais ou eram idênticos à seqüência genômica (*e.g. licA*) ou eram diferentes (*e.g.* P97-like), consistindo sempre em adenosinas **[TAB. 4.2]**.

TABELA 4.2 Análise da extremidade 5' dos transcritos

(continua)

Gene		Seqüência [a]	TSS [b]	Clones [c]
sipS	genoma:	TGTTTTTTTATGATAAAATATCAAAGAAATAAATTTAATAACTATGTTAAA		
	seq nº 1:	GAAATAAATTTAATAACTATGTTAAA	−60	2 / 12
	seq nº 2:	AAAAATAAATTTAATAACTATGTTAAA	−59	10 / 12
	seq nº 3:	AAAATAAATTTAATAACTATGTTAAA		
recA	genoma:	TAAAATGTTTAAAAAATATTAAATTAGAATAAGCAACAAACCAAAACCGAA		
	seq nº 1:	AGAATAAGCAACAAACCAAAACCGAA	−72	8 / 19
	seq nº 2:	AAGCAACAAACCAAAACCGAA	−67	7 / 19
licA	genoma:	TGAAAAATATAATAAAATTTCCAGTATGAAAAAAATTTTGATTGGATTTAC		
	seq nº 1:	ATGAAAAAAATTTTGATTGGATTTAC	+34	6 / 15
	seq nº 2:	AAAATTTTGATTGGATTTAC	+40	9 / 15
uvrC	genoma:	ACCAATTTTTTGTTAAAATTATAATAATTATCAAAAATTTGTTTATATTTT		
	seq nº 1:	AATTATCAAAAATTTGTTTATATTTT	−78	11 / 11
clpB	genoma:	AAAAAATATGTTATAATTTATTTTGTTAAGAAAAATAAAAAATTAAAAAGG		
	seq nº 1:	TTAAGAAAAATAAAAAATTAAAAAGG	−17	10 / 10
rpsJ	genoma:	TATTTTTTGTGTATAATTTAATCTTACCGTTAAAGGATGATGATGCATAAG		
	seq nº 1:	ACCGTTAAAGGATGATGATGCATAAG	−69	9 / 10
P97	genoma:	AAAAAAAAAAAGTATAATTTTAATTGTACAAGTTAAATAAATTTTTCACTT		
	seq nº 1:	GTACAAGTTAAATAAATTTTTCACTT	−138	12 / 12
glyA	genoma:	GTTTTTTTAGTGTATAATGTGTAAAATTCCAATGTATAAGAAAATCAAACT		
	seq nº 1:	ATTCCAATGTATAAGAAAATCAAACT	−6	12 / 13
0225	genoma:	TATTTTTTATGTTAAAATTATAATCGTAGGGTTAAAAATTGATAATTTTTA		
	seq nº 1:	GTAGGGTTAAAAATTGATAATTTTTA	−88	12 / 23
P97−like	genoma:	ATTAAAATATGGTATAATTTTAATTAAATTGTAAATTCGCGGAGGTGAGCT		
	seq nº 1:	AAAAAAAAATTGTAAATTCGCGGAGGTGAGCT		
	seq nº 2:	AAAAAAATTGTAAATTCGCGGAGGTGAGCT	−57	12 / 14
	seq nº 3:	AAAAAATTGTAAATTCGCGGAGGTGAGCT		
	seq nº 4:	AAAATTGTAAATTCGCGGAGGTGAGCT		
0279	genoma:	TTTAAAAATAGTGTAAAATTGTTAAATTATGAGTGTTCTAACCATTTCAAA		
	seq nº 1:	ATTATGAGTGTTCTAACCATTTCAAA	+1	11 / 11
glpK	genoma:	TGGATTAAAATGTTATAATTCAAATATATAAAAATTGAAAGGATTTAAAAA		
	seq nº 1:	ATATAAAAATTGAAAGGATTTAAAAA	−27	12 / 17
	seq nº 2:	TATAAAAATTGAAAGGATTTAAAAA	−26	2 / 17
	seq nº 2:	ATAAAAATTGAAAGGATTTAAAAA	−25	3 / 17
P37	genoma:	CTTCCTTCTATGATATAATAATTTCAGGTAAACAGAGGTGTCCATTTTTGG		
	seq nº 1:	AGGTAAACAGAGGTGTCCATTTTTGG	−144	10 / 11
efp	genoma:	TTTTTTTATGCTATAATTTATAGTTACTTTGTTTATTGTCAAGTGGAGGCG		
	seq nº 1:	ACTTTGTTTATTGTCAAGTGGAGGCG	−28	10 / 19
pgk	genoma:	TCAACTTAATAGTTTATAATATAACAAAAATAAAATTTTCAAAGGATAAAA		
	seq nº 1:	AAAAAAATAAAATTTTCAAAGGATAAAA	−26	10 / 10
	seq nº 2:	AAAAAATAAAATTTTCAAAGGATAAAA		
46K	genoma:	TTGATTTTTATAGTATAATTTATTTGTATAATTGAATTAACTTGATTTGAA		
	seq nº 1:	GTATAATTGAATTAACTTGATTTGAA	−36	10 / 16
	seq nº 2:	ATAATTGAATTAACTTGATTTGAA	−34	4 / 16
gyrA	genoma:	AAAAAATATGGTATAATTTATACTTACCACGCGCTTTTATATCAAAAAGGA		
	seq nº 1:	ACCACGCGCTTTTATATCAAAAAGGA	+14	12 / 15

TABELA 4.2 Análise da extremidade 5' dos transcritos

(conclusão)

Gene		Seqüência [a]	TSS [b]	Clones [c]
pyrH	genoma:	TTTCAAAAATAAAGTATAATAAAAGAAACTTTTAAGGCTAATATGGACTCA		
	seq nº 1:	AAAAAACTTTTAAGGCTAATATGGACTCA		
	seq nº 2:	AAAAACTTTTAAGGCTAATATGGACTCA	−17	8 / 15
	seq nº 3:	AAAACTTTTAAGGCTAATATGGACTCA		
	seq nº 4:	AACTTTTAAGGCTAATATGGACTCA	−16	2 / 15
	seq nº 5:	ACTTTTAAGGCTAATATGGACTCA	−15	4 / 15
ktrA	genoma:	TTCAATTTTGGTCTACAATTTAGTCAAATGAAACGAGCAAATATCTGTATA		
	seq nº 1:	AAAAAATGAAACGAGCAAATATCTGTATA		
	seq nº 2:	AAAAATGAAACGAGCAAATATCTGTATA	−2	13 / 15
	seq nº 3:	AAAATGAAACGAGCAAATATCTGTATA		
rplJ	genoma:	TTTCGCTTTCTGGTATAATTCAAAAACGCAATACATAGAGATAATAACTGC		
	seq nº 1:	ACGCAATACATAGAGATAATAACTGC	−101	8 / 17
	seq nº 2:	GCAATACATAGAGATAATAACTGC	−99	5 / 17
dam	genoma:	CCTATTTGCTTATATAATTTAGTTTATGTCAAAATCTGAATTAAAGCCTTT		
	seq nº 1:	ATGTCAAAATCTGAATTAAAGCCTTT	+1	12 / 17
	seq nº 2:	TGTCAAAATCTGAATTAAAGCCTTT	+2	5 / 17
leuS	genoma:	TAAAAAATTATGCTATAATTTAGGTAACTATCATGTTAGATCACCGAGCTA		
	seq nº 1:	AACTATCATGTTAGATCACCGAGCTA	−7	10 / 11
P146	genoma:	TTAACTTCTATAGTATAATTATTGTATCACTTCGTCTATTATAATAATATA		
	seq nº 1:	ATCACTTCGTCTATTATAATAATATA	−35	12 / 12

[a] Alinhamento mostrando todas as variações encontradas nas seqüências correspondentes à extremidade 5' dos transcritos dos 23 genes. Genoma: seqüência correspondente à região do genoma de *M. hyopneumoniae* 7448 que compreende a extremidade 5' do transcrito. Seq n°: seqüências referentes às diferentes extremidades 5' encontradas. Em laranja: seqüência de adenosinas não coincidentes com a seqüência genômica. Em caixa preta: último nucleotídeo coincidente com a seqüência genômica, sendo o possível sítio de início de transcrição.
[b] Posição do sítio de início de transcrição (TSS) em relação ao códon de iniciação anotado. Números negativos: TSS está a montante do gene. Números positivos: TSS está dentro da região codificadora.
[c] O número de clones seqüenciados é informado para cada gene juntamente com a freqüência do respectivo TSS. Somente TSSs observados em pelo menos dois clones diferentes foram considerados.

Em geral, a seqüência mais longa foi a mais comum entre os clones seqüenciados. Uma ou duas seqüências menores, diferindo não mais que seis nucleotídeos, também foram relativamente freqüentes em oito genes (*sipS*, *recA*, *licA*, *glpK*, 46K, *pyrH, rplJ*, *dam*) **[TAB. 4.2]**. Essa variação poderia indicar a existência de TSSs alternativos ou poderia ter sido originada a partir de transcritos processados que foram co-purificados com os transcritos primários, uma vez que ambos estão presentes nas amostras tratadas com TAP e podem apresentar diferenças mínimas de tamanho. Portanto, o nucleotídeo 5' da maior seqüência de cada gene foi considerado como o TSS correto.

Cinco genes (*sipS*, P97-like, *pgk*, *pyrH* e *ktrA*) apresentaram nucleotídeos adicionais na suas extremidades 5' que não eram esperados **[TAB. 4.2]**. Esses nucleotídeos

extras consistiam de uma a seis adenosinas dentro de uma região homopolimérica composta de, pelo menos, três adenosinas. Nesses casos, o último nucleotídeo 5' correspondente à seqüência genômica foi considerado como o TSS correto.

No total, os TSSs de 23 genes de *M. hyopneumoniae* foram identificados **[TAB. 4.2; TAB 4.3]**. Quatro TSSs foram encontrados dentro das regiões codificadoras de seus respectivos genes: 34 bp em *licA*, 14 pb em *gyrA*, 1 pb em MHP7448_0279 e em *dam*. Para estes genes, o próximo códon de iniciação em fase foi adotado como verdadeiro. As distâncias entre os TSSs e o início das regiões codificadoras variaram de 143 pb em P37 a 1 pb em *ktrA*. Os genes *rplJ* e MHP7448_0198 também apresentaram TSSs distantes, estando localizados, respectivamente, a 100 e 137 pb de distância de seus códons iniciadores. No entanto, os TSSs de *licA*, *glyA*, MHP7448_0279 e *leuS* estavam situados a uma distância menor que 10 pb **[TAB 4.3]**.

Com exceção do gene *clpB*, cujo o TSS identificado foi uma timidina, todos os outros genes apresentavam purinas como TSS. No total, 80% dos transcritos iniciavam com um resíduo de adenosina **[TAB. 4.3]**.

4.4 Identificação dos elementos promotores

Os 23 TSSs determinados foram alinhados e as seqüências localizadas imediatamente a 5' deles foram analisadas em busca de padrões que poderiam corresponder a elementos promotores. A ocorrência de seqüências com alta representatividade, em um determinado segmento de seqüência, foi detectada usando a ferramenta Local-Word-Analysis. Quando motivos de seis nucleotídeos foram procurados, em 21 dos 23 genes foram detectados os padrões TATAAT ou TAAAAT localizados entre 5–9 bases a montante do TSS **[TAB. 4.3]**. Outros padrões foram reconhecidos para os genes restantes, com o auxílio dos programas MEME e Wconsensus: AAAAAT para *recA* e TACAAT para *ktrA*.

TABELA 4.3 Regiões promotoras dos genes de *Mycoplasma hyopneumoniae*

Gene		Região 5'	−16[b]		−10	TSS[c]	Dist[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[†]
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[†]
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[†]
MHP7448_0535	*pyrH*	TTTTTTAATACATTTTTTTTCAAAAA	TAAA	G	TATAAT	AAAAGA	16	ATG
MHP7448_0545	*ktrA*	GATAATTTTAAAAATTTTTTCAATTT	TGGT	C	TACAAT	TTAGTCA	1	ATG
MHP7448_0619	*rplJ*	AAAAAATACTTTTTTTATTTTCGCTT	TCTG	G	TATAAT	TCAAAAA	100	TTG
MHP7448_0622	*dam*	TTTTTTAAATAATTTTATCCCTATTT	GCTT	A	TATAAT	TTAGTTTA	14	TTA[†]
MHP7448_0647	*leuS*	ACTTTTGGCTTTTAATTTTAAAAAAT	TATG	C	TATAAT	TTAGGTA	6	ATG
MHP7448_0663	P146	GAGAAATTTTTTTAATTTTTAACTTC	TATA	G	TATAAT	TATTGTA	34	ATG
			••••	•	••••••	• PSSM 12 col.		
			•• ••••	•	••••••	• PSSM 14 col.		
			••• ••••	•	••••••	•• PSSM 16 col.		

[a] Note que não há evidência do elemento −35 (TTGACA) nesta região.
[b] Região onde o elemento −16 foi encontrado.
[c] Os sítios de início de transcrição estão sublinhados.
[d] Distância (pb) entre o TSS e o códon de iniciação.
[e] Códon de iniciação. [†] Os códons de iniciação dos genes *licA*, 0279, *gyrA* e *dam* foram redefinidos, pois seus TSSs estavam localizados dentro da CDS originalmente anotada.
Caixas pretas: nucleotídeos que ocorrem em mais de 80% dos promotores.
Caixas cinza escuras: nucleotídeos que ocorrem em mais de 70% dos promotores
Caixas cinza claras: guaninas que ocorrem em mais de 40% dos promotores.
Pontos: posições usadas na construção das diferentes PSSMs.

Quatro das seis posições dos diferentes padrões encontrados conservam o mesmo nucleotídeo. Além disso, a timidina é a primeira base em 22 (96%) e a terceira base em 16 (70%) dos 23 sítios identificados. Sendo assim, foi obtida a seqüência consenso TATAAT, a qual é idêntica à seqüência canônica do elemento −10 de um promotor σ^{70} (HAWLEY & MCCLURE, 1983).

Dando continuidade a análise dos sítios promotores, as seqüências foram alinhadas em função dos supostos elementos −10, evidenciando a existência de mais regiões conservadas. Além da base imediatamente a 3' dos hexanucleotídeos ser uma timidina em 73% das seqüências, o padrão TATG foi identificado pela ferramenta Local-Word-Analisys,

uma base a montante dos hexanucleotídeos, em oito genes. **[TAB. 4.3]**. Este motivo corresponde ao consenso 5'-TRTGn-3', uma região −10 estendida comumente encontrada nas bactérias gram-positivas, também conhecida como elemento −16 (VOSKUIL *et al*., 1995; VOSKUIL & CHAMBLISS, 2002). Podemos ressaltar, ainda, que o dinucleotídeo TG – principal determinante dos elementos −10 estendidos, foi encontrado uma base a montante do hexanucleotídeo −10 em outros três genes. Portanto, 11 (48%) das seqüências promotoras continham um provável elemento −10 estendido.

Embora tenha sido possível identificar os prováveis elementos −10 e −16, nenhum padrão conservado, correspondente ao elemento −35, foi encontrado **[TAB. 4.3]**. Interessantemente, uma seqüência periódica AT-rica pôde ser observada quando um *logo* de seqüência foi criado **[FIG. 4.2]**.

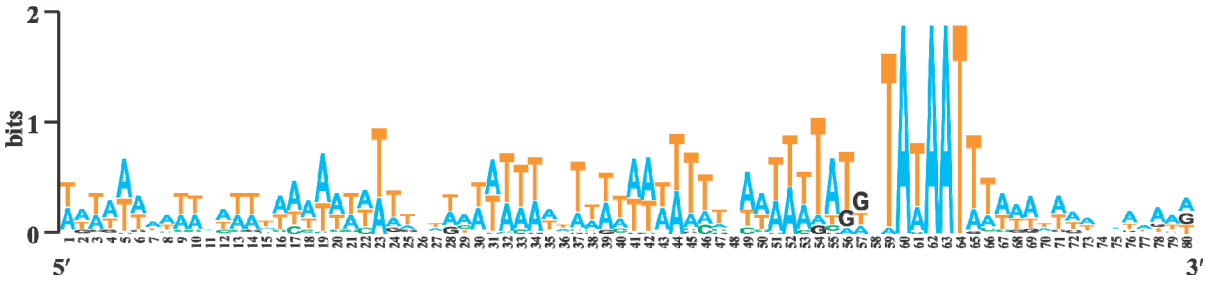


FIGURA 4.2 Conservação das regiões promotoras de *Mycoplasma hyopneumoniae*.
Logo de seqüência, gerado a partir do alinhamento das 23 regiões promotoras definidas, mostrando alta conservação do elemento −10 (posições 59–64), a presença de um elemento −16 semi-conservado (posições 54–57), a ausência de um elemento −35 e um sinal periódico AT-rico distinto que se estende a montante do elemento −10. A região que compreende as posições 54 a 65 foi usada para construir a PSSM de 12 colunas. O eixo vertical mostra a informação contida em *bits*. A altura geral da pilha de nucleotídeos indica a conservação da seqüência naquela posição, e a altura de cada nucleotídeo em cada pilha indica a sua freqüência relativa nessa posição.

4.5 Comparação entre os promotores σ^{70} de diferentes bactérias

As seqüências promotoras de *M. hyopneumoniae* foram comparadas com promotores σ^{70} de diferentes bactérias. Para que esta comparação fosse realizada, os

alinhamentos de regiões promotoras experimentalmente caracterizadas de outras sete espécies, com conteúdo genômico de G+C variado, foram usados para criar *logos*, os quais visualmente representam a conservação de seqüências **[FIG. 4.3]**.

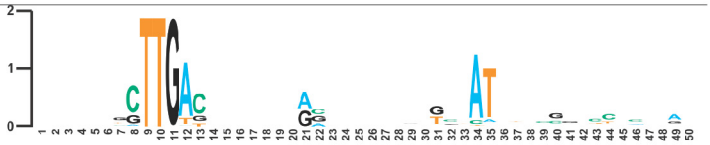
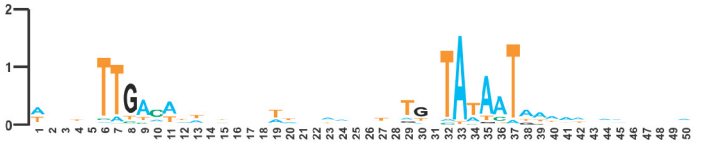
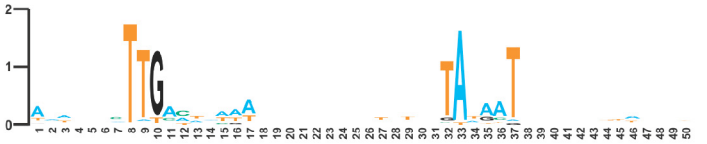
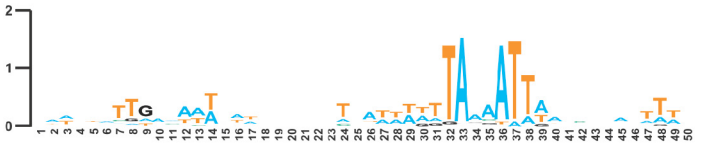
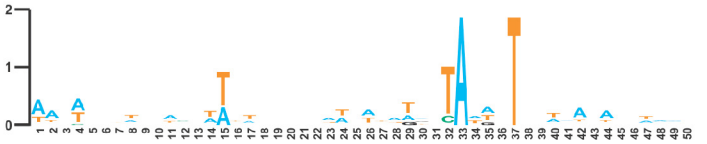
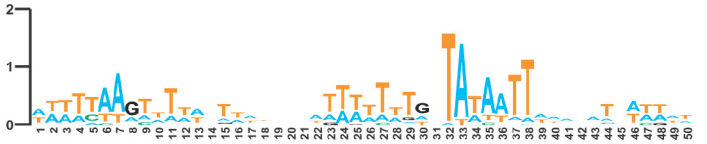
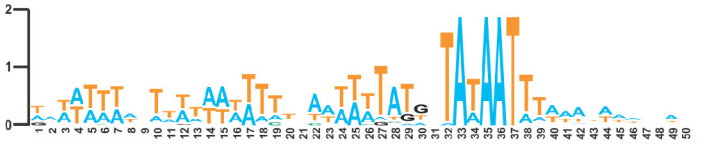
Bacteria	Genoma (kb)	G+C (%)	Logo de Seqüência (5'–3')
S. meliloti	3.654	62,7	
E. coli	5.498	50,5	
B. subtilis	4.215	43,5	
C. trachomatis	1.042	41,3	
M. pneumoniae	816	40	
P. marinus	1.657	30,8	
C. jejuni	1.616	30,6	
M. hyopneumoniae	920	28,5	

FIGURA 4.3 Regiões promotoras de σ^{70} de diferentes espécies bacterianas.
Logos de seqüências mostrando a perda de conservação do elemento −35 à medida que o conteúdo genômico de G+C decresce. Os seguintes números de seqüências promotoras foram usados para gerar os *logos*: 25 sítios de *S. meliloti*, 59 sítios de *E. coli*, 142 sítios de *B. subtilis*, 41 sítios de *C. trachomatis*, 35 sítios de *M. pneumoniae*, 25 sítios de *P. marinus*, 21 sítios de *C. jejuni* e 23 sítios de *M. hyopneumoniae*. O eixo vertical mostra a informação contida em *bits*. A altura geral da pilha de nucleotídeos indica a conservação da seqüência naquela posição, e a altura de cada nucleotídeo em cada pilha indica a sua freqüência relativa nessa posição.

Os resultados presentes na **FIG. 4.3** sugerem que a ocorrência de elementos −35 em promotores σ^{70} estaria relacionada com o conteúdo de G+C do organismo. Os promotores das espécies *Sinorhizobium meliloti*, *Escherichia coli*, *Bacillus subtilis, Chlamydia trachomatis* e *Mycoplasma pneumoniae*, cujo conteúdo genômico de G+C era ≥40%, apresentam o trinucleotídeo TTG, referente ao elemento −35, enquanto os promotores de *Prochlorococcus marinus*, *Campylobacter jejuni* e *M. hyopneumoniae*, com o conteúdo genômico de G+C ≤30,8%, não possuem esse trinucleotídeo conservado.

A comparação dos *logos* de seqüência também mostrou que o elemento −10 é mais conservado em *M. hyopneumoniae* do que nas outras espécies bacterianas. Uma observação importante foi a de que esse elemento era precedido por um dinucleotídeo TG, característica compartilhada com *B. subtilis* e *C. jejuni*, e que indica a existência de um elemento −10 estendido. Outra particularidade evidenciada foi a presença de seqüências periódicas AT-ricas localizadas a montante do elemento −10 de *M. hyopneumoniae* e *C. jejuni*.

4.6 Conservação das regiões de ligação ao promotor dos fatores σ^{70} de diferentes bactérias

Além dos promotores de *S. meliloti*, *E. coli*, *B. subtilis*, *C. trachomatis*, *M. pneumoniae*, *P. marinus*, *C. jejuni* e *M. hyopneumoniae*, os fatores σ^{70} destas bactérias também foram analisados. Esta análise teve como objetivo verificar se as regiões de sigma, responsáveis pelo reconhecimento dos elementos promotores, refletiam as variações encontradas entre os promotores.

A partir do alinhamento dos fatores σ^{70} das diferentes espécies, foram identificadas as regiões 2.4, 3.0 e 4.2 **[FIG. 4.4]**, as quais interagem de forma direta com elementos promotores −10, −16 e −35, respectivamente.

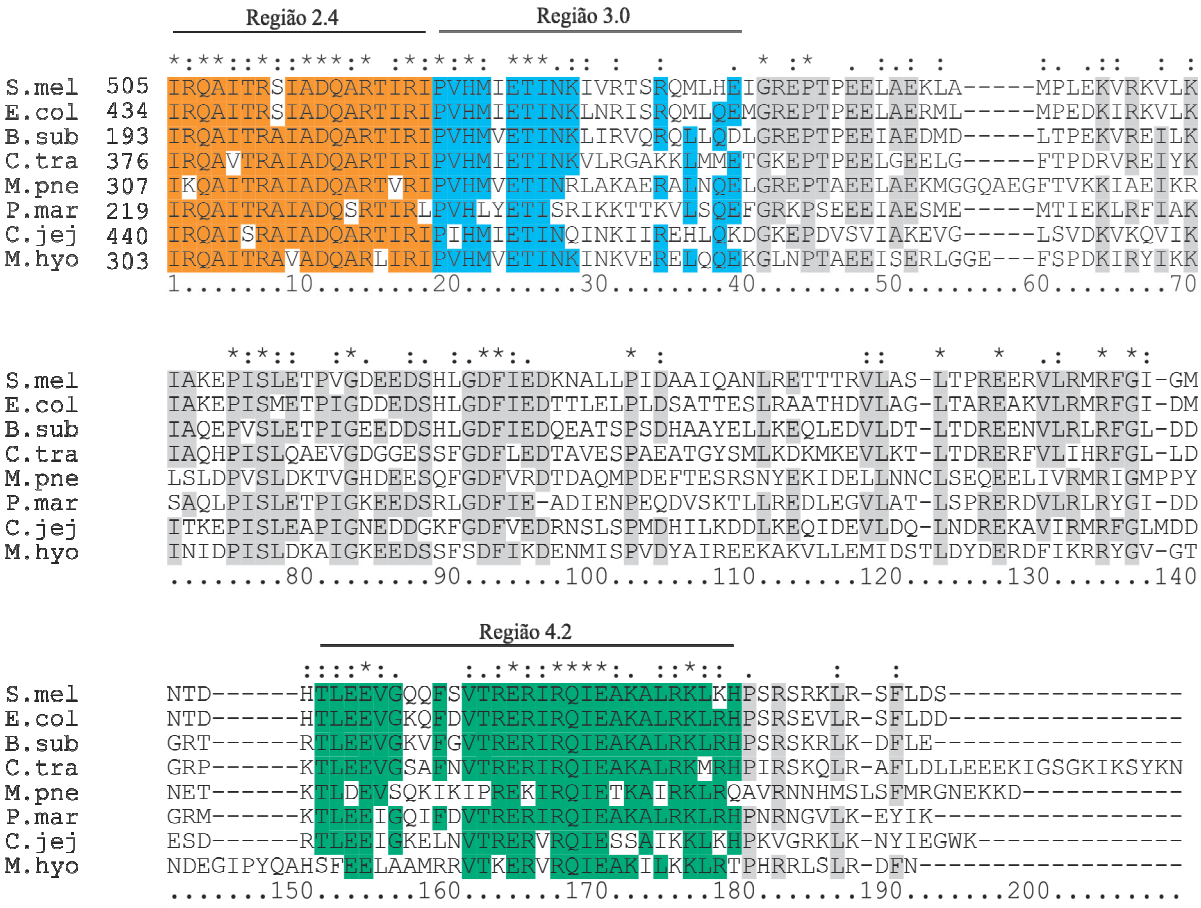


FIGURA 4.4 Alinhamento das seqüências de aminoácidos dos fatores σ^{70} de diferentes espécies bacterianas.
Alinhamento mostrando a extremidade carboxi-terminal dos fatores σ^{70}, onde estão localizadas as regiões 2.4, 3.0 e 4.2, que são responsáveis pela interação direta do fator σ^{70} com os elementos promotores −10, −16 e −35, respectivamente. Em destaque estão as regiões cuja variação entre as seqüências foi mensurada: região 2.4 (posições 1−18), região 3.0 (posições 19−39) e região 4.2 (posições 151−179). Os fundos em laranja, azul, cinza ou verde indicam mais de 60% de identidade entre as seqüências. S.mel, *S. meliloti*; E.col, *E. coli*; B.sub, *B. subtilis*; C.tra, *C. trachomatis*; M.pne, *M. pneumoniae*; P.mar, *P. marinus*; C.jej, *C. jejuni*; M.hyo, *M. hyopneumoniae*.

A variação existente entre as seqüências, em cada uma das três regiões de σ^{70}, foi mensurada através do cálculo da distância evolutiva, a qual reflete ao número de substituições ocorrido ao longo do tempo, permitindo avaliar a conservação de uma seqüência em relação à outra. Essa comparação foi feita tendo como referência as regiões 2.4, 3.0 e 4.2 do fator σ^{70} de

E. coli, uma vez que as interações dos aminoácidos dessa proteína com o promotor estão bem caracterizadas. Sendo assim, quanto maior o valor encontrado, maior a diferença entre as seqüências de *E. coli* e as seqüências da espécie em questão **[TAB. 4.4]**.

TABELA 4.4 Conservação das regiões 2.4, 3.0 e 4.2 entre os fatores σ^{70} das diferentes espécies bacterianas

Bactéria	Região 2.4	Região 3.0	Região 4.2
E. coli	0	0	0
S. meliloti	0	0,205	0,11
B. subtilis	0,048	0,313	0,074
C. trachomatis	0,097	0,594	0,149
M. pneumoniae	0,147	0,621	0,518
P. marinus	0,152	0,856	0,116
C. jejuni	0,097	0,446	0,388
M. hyopneumoniae	0,162	0,429	0,701

Nota: A conservação das seqüências de cada região foi verificada em comparação as regiões 2.4, 3.0 e 4.2 de *E. coli*. Os valores refletem o número de substituições de aminoácidos existentes entre a seqüência de *E. coli* e a seqüência da outra espécie em questão; quanto maior o valor, maior a diferença entre elas.

A região 2.4, que reconhece o elemento −10, apresenta as menores taxas de substituição de aminoácidos, indicando que essa é, dentre as três, a região mais conservada **[TAB. 4.4]**. Este resultado está em acordo com o fato de que o elemento −10 é o único que pôde ser identificado nos promotores de todas as oito espécies de bactérias analisadas **[FIG. 4.3]**. Da mesma forma, a menor conservação das regiões de σ^{70} que interagem com os elementos −16 e −35 condiz com a observação de que esses elementos não ocorrem em todos os promotores **[FIG. 4.3]**.

C. trachomatis, *M. pneumoniae* e *P. marinus*, os quais apresentam as maiores distâncias evolutivas em relação à região 3.0 **[TAB. 4.4]**, são os mesmos organismos cujos consensos promotores não possuem indícios de uma região −10 estendida (elemento −16) **[FIG. 4.3]**.

Com exceção de *P. marinus*, as demais espécies em que o elemento −35 não ocorre ou é pouco evidente **[FIG. 4.3]**, também apresentam uma região 4.2 bastante divergente

da região 4.2 de *E. coli* **[TAB. 4.4]**, a qual faz parte das bactérias cujos promotores comumente possuem um elemento −35 bem determinado.

4.7 Construção de uma PSSM para predição de promotores de *M. hyopneumoniae*

O alinhamento manual dos 23 promotores definidos de *M. hyopneumoniae* foi usado para construir a PSSM de 12 colunas **[TAB. 4.3; FIG. 4.2]**. Com o objetivo de validar este alinhamento e as posições incluídas na matriz, outras duas matrizes foram independentemente construídas utilizando os programas MEME e Wconsensus. Ambos os algoritmos empregaram as mesmas 12 posições, usadas na matriz inicial, para construir suas matrizes. Entretanto, estas últimas incluíram algumas posições a mais, gerando PSSMs de 14 e 16 colunas **[TAB. 4.3]**.

Após a obtenção das três matrizes, elas foram reconstruídas, excluindo os sítios repetidos, a fim de minimizar o viés gerado pelos motivos mais freqüentes. Sendo assim, a PSSM de 12 colunas **[TAB. 4.5]** foi composta por 20 dos 23 sítios determinados, e as PSSMs de 14 e 16 colunas **[TAB. 4.6; TAB. 4.7]** foram compostas por 22 deles.

TABELA 4.5 PSSM de 12 colunas baseada nos promotores experimentalmente definidos de *Mycoplasma hyopneumoniae*

	−17	**−16**	**−15**	**−14**	**−13**	**−12**	**−11**	**−10**	**−9**	**−8**	**−7**	**−6**
A	2	14	2	4	5	1	20	6	20	20	0	5
C	0	2	0	0	2	0	0	1	0	0	0	0
G	2	1	5	9	8	0	0	0	0	0	0	1
T	16	3	13	7	5	19	0	13	0	0	20	14

Nota: o alinhamento e as posições utilizadas para construir esta matriz constam na **TAB. 4.3**.

TABELA 4.6 PSSM de 14 colunas baseada nos promotores experimentalmente definidos de *Mycoplasma hyopneumoniae*

	−19	−18	−17	−16	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6
A	10	8	2	16	2	5	5	1	22	6	22	22	0	5
C	0	2	0	2	0	0	3	0	0	1	0	0	0	0
G	0	0	2	1	5	10	9	0	0	0	0	0	0	1
T	12	12	18	3	15	7	5	21	0	15	0	0	22	16

Nota: o alinhamento e as posições utilizadas para construir esta matriz constam na **TAB. 4.3**.

TABELA 4.7 PSSM de 16 colunas baseada nos promotores experimentalmente definidos de *Mycoplasma*. *hyopneumoniae*

	−20	−19	−18	−17	−16	−15	−14	−13	−12	−11	−10	−9	−8	−7	−6	−5
A	8	10	8	2	16	2	5	5	1	22	6	22	22	0	5	5
C	1	0	2	0	2	0	0	3	0	0	1	0	0	0	0	2
G	0	0	0	2	1	5	10	9	0	0	0	0	0	0	1	1
T	13	12	12	18	3	15	7	5	21	0	15	0	0	22	16	14

Nota: o alinhamento e as posições utilizadas para construir esta matriz constam na **TAB. 4.3**.

Posteriormente, a capacidade preditiva das matrizes foi avaliada e comparada, para que fosse possível identificar a que apresentava o melhor desempenho. Para fazer essa avaliação foi usada a ferramenta Matrix-Quality. Este programa baseia-se em uma análise combinada de distribuições de valores de peso teóricos e empíricos para estimar a capacidade da PSSM em distinguir os possíveis sítios de ligação de fatores de transcrição a partir do *background* genômico (MEDINA-RIVERA *et al*., 2011).

A distribuição teórica corresponde à distribuição que seria esperada para os valores de peso obtidos pela matriz ao longo de uma seqüência randômica de tamanho infinito, gerada de acordo com o modelo de *background*. Desta forma, para cada valor de peso é atribuído um valor P, o qual indica a probabilidade de se observar, ao acaso, um sítio com pontuação igual ou maior a dado valor de peso. Conseqüentemente, a distribuição teórica fornece uma estimativa da taxa de falso-positivos (*false positive rate*; FPR) associada a cada valor de peso possível.

A distribuição empírica corresponde à distribuição observada para os valores de peso obtidos pela matriz ao longo de seqüências de interesse (*e.g*. seqüencias intergênicas a montante dos genes). Essas seqüências são compostas, predominantemente, por seqüências que não são sítios de ligação de fatores de transcrição (*background* genômico), intercaladas com alguns sítios biologicamente funcionais (MEDINA-RIVERA *et al*., 2011)

Para cada matriz, foram calculadas cinco distribuições empíricas e teóricas diferentes **[FIG. 4.5]**, com o objetivo de escolher o melhor modelo de *background*. Além, de esse parâmetro ser importante por determinar como será a seqüência randômica utilizada no cálculo da distribuição teórica, ele também influencia o cálculo dos valores de peso e, conseqüentemente, afeta a performance das matrizes.

Todas as distribuições empíricas foram obtidas a partir do mesmo conjunto de seqüências de interesse, o qual foi composto por até 250 bases a montante e 50 bases a jusante do códon de iniciação de todas as CDSs de *M*. *hyopneumoniae* (as bases a jusante também foram investigadas, pois alguns TSSs foram encontrados dentro de seus genes). As distribuições teóricas foram obtidas a partir de diferentes seqüências randômicas geradas de acordo com o modelo de *background* empregado. Foram testadas cinco ordens de Markov (de 0 a 4) como modelo de *background*, utilizando como base a composição nucleotídica de seqüências não-codificantes localizadas a montante de todos os genes de *M*. *hyopneumoniae*.

A **FIG. 4.5** demonstra que a distribuição empírica sobrepõe-se a distribuição teórica nos valores de peso menores (mostrando que nesses valores não é possível distinguir entre possíveis promotores e o *background* genômico), mas se separa nos valores de peso maiores, provavelmente correspondendo à identificação sítios de ligação funcionais. Essa separação entre as extremidades de ambas as distribuições indica a capacidade da matriz em discriminar seqüências que são prováveis sítios de ligação do fator σ de seqüências que não são sítios de ligação (*background* genômico).

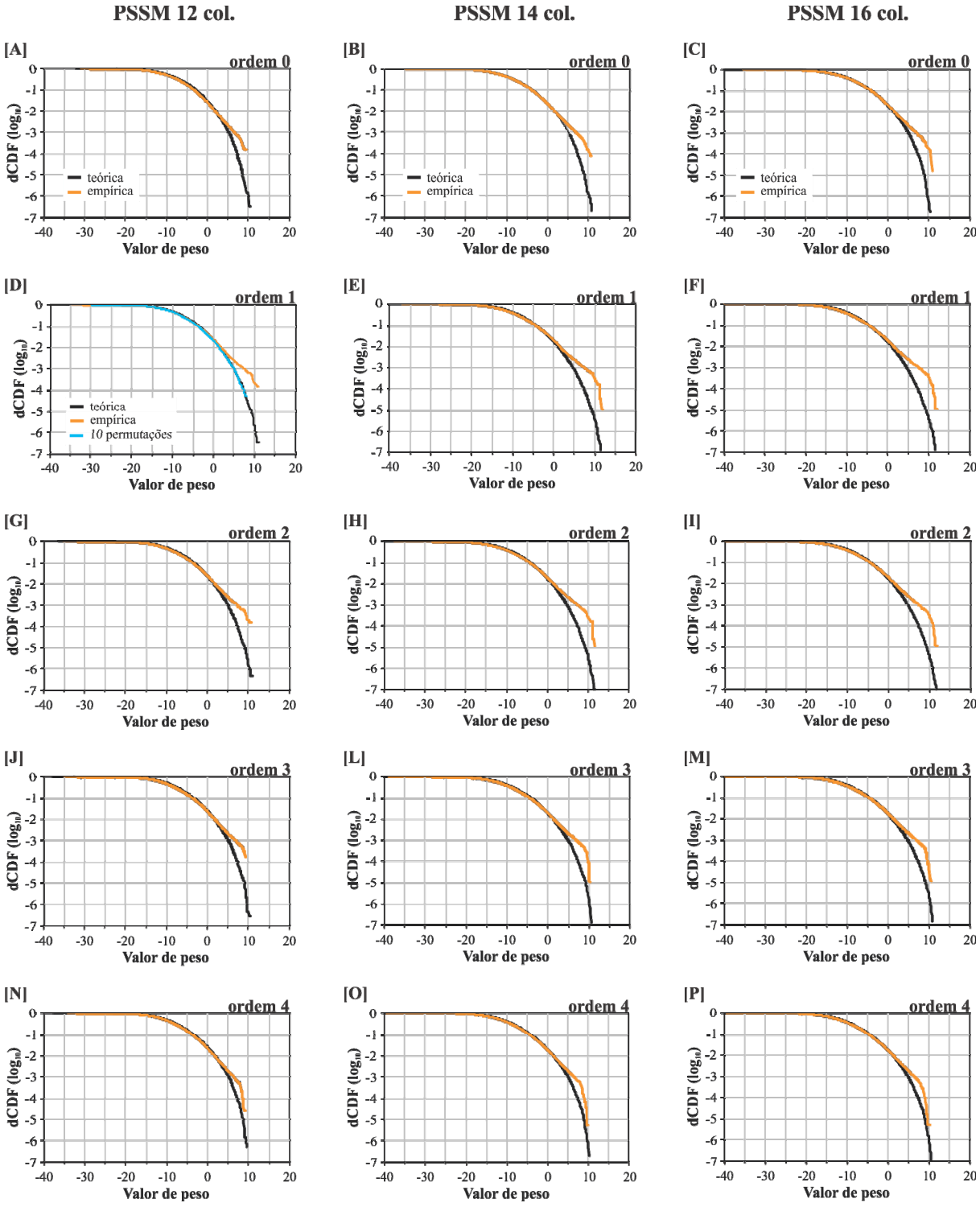


FIGURA 4.5 Distribuições de valores de peso calculadas com as PSSMs de 12, 14 e 16 colunas empregando diferentes modelos de *background*.
Gráficos mostrando as curvas de distribuição teórica (linha preta) e empírica (linha laranja) obtidas com cada PSSM utilizando diferentes ordens de Markov (0 a 4) como modelo de *background*. Note que a distribuição empírica sobrepõe-se a distribuição teórica nos valores de peso menores, e separa-se nos valores maiores.
O gráfico da PSSM de 12 colunas associada a ordem 1 de Markov, também apresenta a média das distribuições dos valores de peso obtidas com 10 matrizes de colunas permutadas (linha azul), a qual sobrepõe-se a distribuição teórica, confirmando que esta última pode ser considerada uma estimativa apropriada da FPR.
A função de distribuição cumulativa decrescente (dCDF, ordenada) indica o valor P, ou seja, a probabilidade de observar um sítio com valor de peso maior ou igual a um determinado valor de peso (abscissa).

As diferenças entre as duas distribuições foram calculadas para comparar o desempenho das matrizes quando associadas com os diferentes modelos de *background*, com o objetivo de determinar a melhor combinação. Isso foi feito através do cálculo da diferença de valor de peso normalizada (*normalized weight score difference*, NWD). Nesta análise, para cada valor de freqüência, é calculada a diferença do valor de peso (*weight score difference*, WD), a qual é definida como a diferença entre os valores de peso observados na distribuição empírica e os valores de peso esperados na distribuição teórica para um dado valor P. Ou seja, o WD pode ser visto como a distância horizontal entre as curvas de distribuição. Como matrizes maiores permitem valores mais altos, o WD é dividido pelo número de colunas da matriz para obter o NWD, o qual possibilita que o desempenho de matrizes de diferentes tamanhos seja comparado. Dessa forma, é possível dizer que quanto maior o NWD, maior a diferença entre as distribuições e, portanto, maior a capacidade de diferenciar sítios funcionais a partir do *background* genômico.

Primeiramente, foram comparadas as curvas de NWD obtidas a partir de uma mesma matriz, porém associadas com diferentes modelos de *background*. Todas as PSSMs apresentaram uma melhor performance com a ordem 1 de Markov **[FIG. 4.6]**.

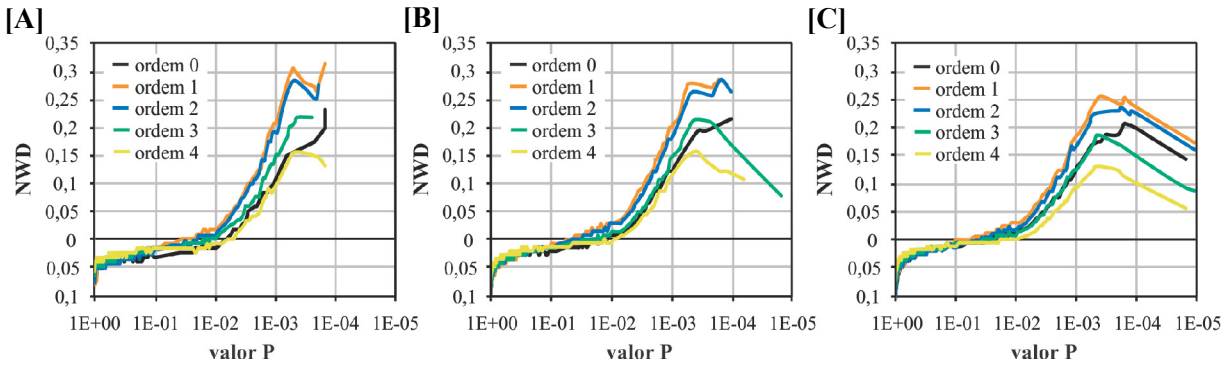


FIGURA 4.6 Desempenho das PSSMs de 12, 14 e 16 colunas usando diferentes ordens de Markov como modelo de *background*.
Cada curva mostra a diferença de valores de peso normalizada (NWD) calculada a partir das distribuições teórica e empírica, obtidas para cada matriz combinadas com cada ordem de Markov (de 0 a 4). Quanto maior o valor de NWD, melhor é a capacidade preditiva da matriz em distinguir prováveis sítios promotores a partir das seqüências de interesse. (A) PSSM de 12 colunas. (B) PSSM de 14 colunas. (C) PSSM de 16 colunas.

A comparação entre as matrizes, usando como modelo de *background* a ordem 1 de Markov, mostrou que a PSSM de 12 colunas obteve os maiores valores de NWD **[FIG. 4.7]**, indicando que esta matriz possui uma maior capacidade preditiva.

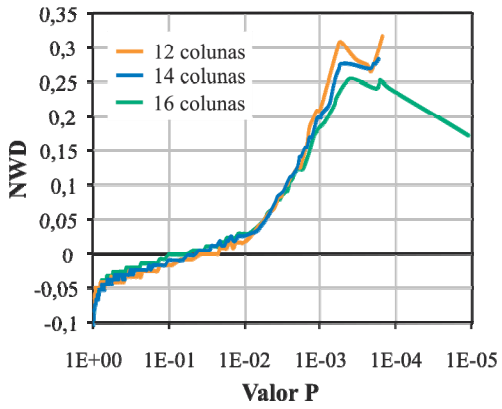


FIGURA 4.7 Desempenho das PSSMs de 12, 14 e 16 colunas usando a ordem 1 de Markov como modelo de *background*.
Cada curva mostra a diferença de valores de peso normalizada (NWD) calculada a partir das distribuições teórica e empírica, obtidas para cada matriz combinadas com a ordem de Markov 1. Quanto maior o valor de NWD, melhor é a capacidade preditiva da matriz em distinguir prováveis sítios promotores a partir das seqüências de interesse.

A análise destes resultados permitiu a definição da PSSM e do modelo de *background*, conduzindo à realização de um controle complementar. Embora a distribuição teórica já seja considerada um controle, uma vez que indica a proporção de falso-positivos esperada, a seqüência randômica, a partir da qual ela é calculada, pode não representar um modelo biológico realista. Portanto, o controle ideal seria aquele obtido a partir de um conjunto de seqüências, ao qual, sabidamente, o fator de transcrição em questão não se liga. Entretanto, evidências experimentais deste tipo não estão disponíveis para *M. hyopneumoniae*.

Um controle alternativo pode ser feito verificando o conjunto de seqüências de interesse com matrizes randomizadas, as quais são geradas através da permutação das colunas da PSSM original. As matrizes de colunas permutadas têm a vantagem de preservar características importantes da matriz, tais como a composição de nucleotídeos e o número de sítios. A distribuição de valores de peso obtida dessa forma pode ser considerada uma

"estimativa empírica da FPR" (MEDINA-RIVERA *et al*., 2011).

Para fazer esse controle, o mesmo conjunto de seqüências, usado para o cálculo da distribuição empírica, foi examinado por dez matrizes de colunas permutadas, derivadas a partir da PSSM de 12 colunas. Nessa análise foram obtidas dez curvas de distribuição de valores de peso, cuja média sobrepôs-se a distribuição teórica por toda sua extensão, sem haver qualquer separação nos valores de peso mais altos **[FIG. 4.5 D]**. Este resultado confirma que a distribuição teórica pode ser considerada uma estimativa apropriada da FPR, e que a divergência observada na distribuição empírica, obtida com a PSSM de 12 colunas original, corresponde a prováveis sítios promotores especificamente detectados pela matriz no genoma (MEDINA-RIVERA *et al*., 2011).

4.8 Determinação do valor de corte

O gráfico de distribuições de valores de peso da PSSM de 12 colunas mostra que as curvas teórica e empírica começam a se separar próximo ao valor de peso 3 **[FIG. 4.8]**, o que provavelmente corresponde à identificação de sítios promotores funcionais.

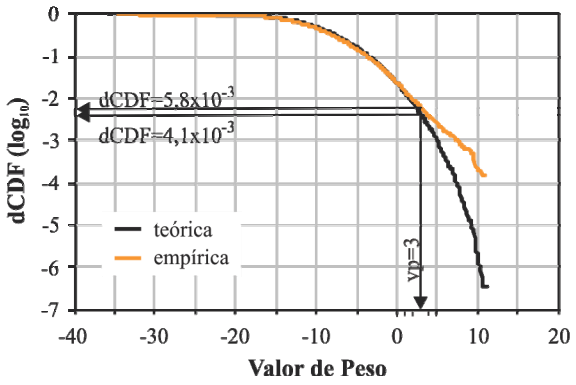


FIGURA 4.8 Distribuições dos valores de peso da PSSM de 12 colunas.
Gráfico mostrando as curvas de distribuição teórica (linha preta) e empírica (linha laranja), obtidas com a PSSM de 12 colunas associada com a ordem 1 de Markov como modelo de *background*. Note que as curvas começam a se separar no valor de peso de 3, indicando que os promotores estão sendo distinguidos a partir do *background* genômico. A função de distribuição cumulativa decrescente (dCDF, ordenada) indica o valor P, ou seja, a probabilidade de observar um sítio com pontuação igual ou maior a um determinado valor de peso (abscissa).

No valor de peso 3, a função de distribuição cumulativa decrescente (*decreasing cumulative distribution function*, dCDF – informa o valor P) na distribuição teórica foi de 4,1 × 10^{-3}, e na distribuição empírica foi de 5,8 × 10^{-3} **[FIG. 4.8]**. Isso significa que para aproximadamente 6 sítios encontrados nas seqüências a montante dos genes, pode-se esperar que aproximadamente 4 sítios sejam falso-positivos. Portanto, a incidência de falso-positivos em relação à freqüência observada de sítios nas seqüências de interesse era demasiadamente alta neste ponto. Entretanto, a partir do valor de peso 3 em diante, a diferença entre as curvas de distribuições empírica e teórica, ou seja, entre as freqüências observada e esperada, aumentou gradualmente **[FIG. 4.8]**. Conseqüentemente, a escolha de um valor limiar, que permitisse a identificação abrangente dos promotores associada a uma FPR relativamente baixa, fez-se necessária.

A definição do valor de corte foi realizada através de uma abordagem complementar descrita por Cases *et al*. (2003). Essa metodologia compara as distribuições de valores de peso de promotores preditos nas fitas senso das seqüências intergênicas (mesma fita do gene a jusante – orientação "correta"), com as de promotores preditos nas fitas anti-senso das seqüências intergênicas (fita oposta ao gene a jusante – orientação "incorreta"). A proposta desta análise considera que os falso-positivos devem estar homogeneamente distribuídos em ambas as fitas, enquanto os verdadeiro-positivos devem estar orientados corretamente.

As seqüências de orientação "correta" consistiram no mesmo conjunto de seqüências de interesse empregado no cálculo da distribuição empírica, com exceção daquelas localizadas entre genes divergentes, uma vez que poderiam conter promotores em ambas as orientações. Entretanto, as seqüências de orientação "incorreta" consistiram no complemento reverso das de orientação "correta". A ocorrência de possíveis promotores, nestas seqüências,

foi determinada pelo programa Matrix-Scan usando a PSSM de 12 colunas e a ordem 1 de Markov como modelo de *background*.

A proporção de promotores com orientação "incorreta" em relação aos promotores com orientação "correta" foi alta nas primeiras três faixas de valores de peso verificadas **[FIG. 4.9]**, indicando uma alta ocorrência de falso-positivos. Porém, a partir do valor de peso de 6,5, a incidência de promotores orientados incorretamente diminuiu consideravelmente, sugerindo que este seria um valor de corte razoável, sendo, portanto, adotado como tal. Para este valor de peso eram esperados 2,4 falso-positivos (dCDF = $2,4 \times 10^{-4}$) em cada 14,2 promotores preditos (dCDF = $1,42 \times 10^{-3}$).

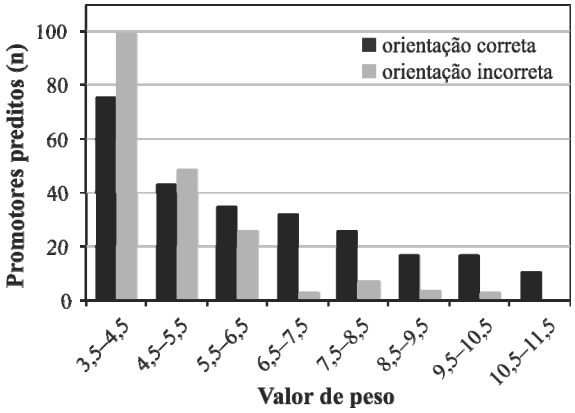


FIGURA 4.9 Definição do valor de corte.
Distribuição dos valores de peso de promotores com orientação correta e incorreta, preditos nas regiões intergênicas de *M. hyopneumoniae*. Note que a partir do valor de peso de 6,5, as freqüências de promotores orientados de forma incorreta são muito menores que as freqüências dos promotores com orientação correta.

A relação entre a FPR e a sensibilidade do valor de corte foi verificada através da curva ROC gerada pelo programa Matrix-Quality **[FIG. 4.10 B]**. A sensibilidade de uma matriz é estimada com base na fração de sítios experimentalmente definidos que recebem uma pontuação acima do valor de corte determinado (MEDINA-RIVERA *et al*., 2011). Então, para avaliar a sensibilidade da PSSM de 12 colunas, essa foi utilizada para determinar o valor de peso dos 20 promotores definidos que foram utilizados para construí-la. Possíveis vieses

gerados nessa estimativa foram corrigidos através do procedimento de *Leave-One-Out* (LOO). Nesse processo, a PSSM de 12 colunas é reconstruída sem o sítio a ser avaliado, e, só então, é utilizada para atribuir o valor de peso ao sítio LOO, ou seja, ao sítio "deixado de fora".

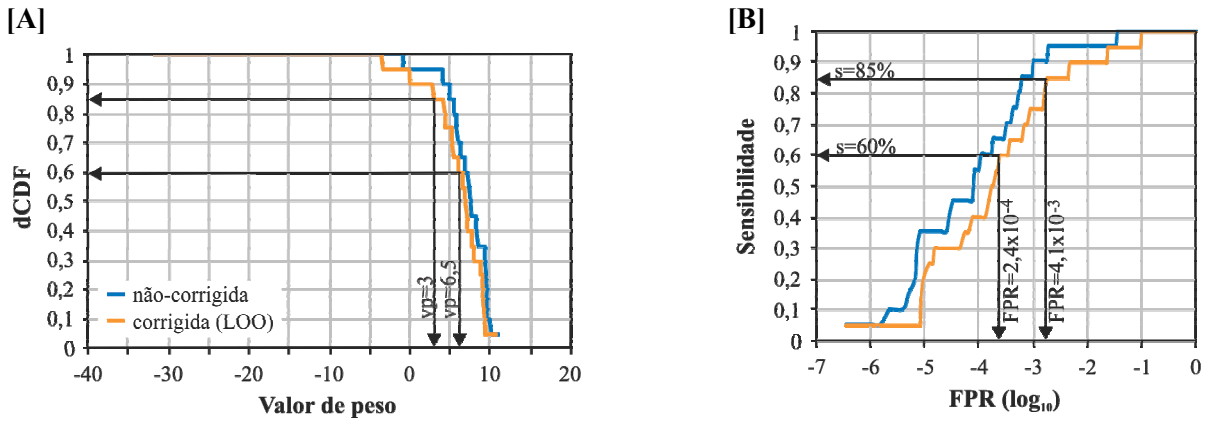


FIGURA 4.10 Relação entre a sensibilidade e a FPR da PSSM de 12 colunas.
[A] Distribuição dos valores de peso dos sítios definidos e que foram usados para construir a matriz. Azul – valores de peso não-corrigidos atribuídos pela matriz aos sítios definidos. Laranja – valores de peso corrigidos usando o procedimento *Leave-One-Out* (LOO). A ordenada indica a probabilidade de observar um sítio com valor de peso maior ou igual a um dado valor de peso (abscissa). [B] Curva ROC indicando o risco de falso-positivos associado com uma sensibilidade específica. Ambos os gráficos mostram a diferença entre as estimativas não-corrigida (azul) e corrigida (laranja). O dCDF (ordenada) indica a sensibilidade (fração de sítios detectados) e a abscissa mostra o FPR correspondente. Note que dCDF da fig. A corresponde à sensibilidade na fig. B.

A sensibilidade estimada para o valor de peso 3, que é referente ao ponto em que as curvas de distribuição teórica e empírica começavam a se separar, foi de 85% **[FIG. 4.10]**, uma vez que, no procedimento LOO, 17 dos 20 sítios promotores definidos receberam pontuação maior que este valor de peso. Porém, embora a sensibilidade tenha sido alta para o valor de peso 3, a FPR associada a ele também era, correspondendo a $4{,}1 \times 10^{-3}$ **[FIG. 4.8; FIG. 4.10]**. Entretanto, o valor de corte escolhido (vp = 6,5), cuja FPR era de $2{,}4 \times 10^{-4}$, apresentou uma sensibilidade estimada de 60% **[FIG. 4.10]**. Sendo assim, seria esperado que 60% dos sítios promotores funcionais fossem detectados pela matriz, e que fossem encontrados 2,4 falso-positivos a cada 10.000 bases analisadas, ao usarmos um valor de corte de 6,5.

É interessante notar que a distância existente entre a curva corrigida (LOO) e a curva não-corrigida é mínima, de modo que o viés pôde ser considerado insignificante, o que indica que a matriz não possui problemas dessa ordem.

4.9 Promotores preditos em *M. hyopneumoniae*

Posteriormente a definição dos parâmetros ótimos para a utilização da PSSM de 12 colunas, as seqüências a montante de todas as 657 CDSs de *M. hyopneumoniae* foram analisadas na busca de prováveis promotores através do programa Matrix-Scan. A **TAB. 4.8** apresenta os resultados gerais desta análise.

TABELA 4.8 Predição de promotores de *Mycoplasma hyopneumoniae*

Características genômicas	**Nº**			
CDSs anotadas no genoma	657			
CDSs que possuem uma região a montante < 15 pb	201	/	657	(31%)
CDSs que possuem um gene a montante em orientação divergente	142	/	657	(22%)
CDSs que possuem um gene a montante com a mesma orientação	515	/	657	(78%)
Características dos promotores preditos (valor de peso ≥ 6,5)	**Nº**			
Promotores	201			
CDSs que possuem ao menos um promotor	169	/	657	(26%)
CDSs que possuem:				
um promotor	143	/	169	(84%)
dois promotores	22	/	169	(13%)
três promotores	3	/	169	(2%)
cinco promotores	1	/	169	(< 1%)
CDSs que possuem um gene a montante com orientação divergente e têm ao menos um promotor	76	/	142	(54%)
CDSs que possuem um gene a montante com mesma orientação e têm ao menos um promotor	93	/	515	(18%)
Características dos promotores preditos (valor de peso ≥ 4,2)	**Nº**			
Promotores	409			
CDSs que possuem ao menos um promotor	273	/	657	(42%)
CDSs que possuem um gene a montante com orientação divergente e têm ao menos um promotor	113	/	142	(80%)
CDSs que possuem um gene a montante com mesma orientação e têm ao menos um promotor	160	/	515	(31%)

A utilização do valor de corte de 6,5 permitiu a identificação de 201 sítios promotores a montante de 169 genes diferentes, correspondendo a 26% do total de CDSs. Para a grande maioria dessas CDS (84%) foi encontrado um único promotor putativo, enquanto que 16% apresentaram sítios promotores adicionais.

A **TAB. 4.8** demonstra que foi detectado, ao menos, um promotor em 54% das CDS que possuíam um gene a montante com orientação divergente e em 18% das CDS que tinham um gene a montante com a mesma orientação. No entanto, estas proporções seriam 80% e 31%, respectivamente, se o valor de corte fosse fixado em 4,2, o menor valor de peso obtido para os promotores experimentalmente definidos.

Analisando especificamente os promotores experimentalmente determinados para os genes cujos TSSs foram identificados **[TAB. 4.9]**, é possível observar que 16 dos 23 sítios definidos receberam valores de peso entre 6,9 e 11, seis receberam valores entre 4,2 e 6,3, e um, o promotor de *recA*, pontuou abaixo de zero. A maioria dos sítios promotores definidos correspondia à seqüência predita de maior valor, mas nos genes *uvrC*, P97 e *ktrA*, corresponderam às segundas melhores seqüências preditas (embora nenhuma delas tenha pontuado acima de 6,5).

Três genes, MHP7448_0279, *rplJ* e P146, apresentaram mais de um promotor predito com pontuação maior ou igual ao valor de corte de 6,5 **[TAB. 4.9]**. Em todos os casos, os promotores adicionais estão localizados em regiões distantes do TSS. Essas seqüências podem corresponder a falso-positivos, indicar a existência de transcritos alternativos para um mesmo gene, ou serem promotores de RNAs não-codificantes.

TABELA 4.9 Promotores preditos para os genes que tiveram os TSSs identificados

Gene	Pos. [a†]	Promotor [b†]	VP [c†]	Gene	Pos. [a†]	Promotor [b†]	VP [c†]
sipS	**−6**	**TATGATAAAATA**	**7,5**	P37	**−6**	**TATGATATAATA**	**9,5**
	+29	TATGTTAAAATC	4,3		−74	TCTTATAAAATA	4,9
*recA**	–	–	–	*efp*	**−8**	**TATGCTATAATT**	**9,5**
licA	**−7**	**TATAATAAAATT**	**5,7**	*pgk*	**−5**	**TAGTTTATAATA**	**7,3**
	−26	TATTTTAAAATT	5,1		−151	AATTTTATAATT	4,5
uvrC	+56	TATGGAAAAATT	5,6	46K	**−6**	**TATAGTATAATT**	**9,7**
	−6	**TTTGTTAAAATT**	**4,9**		+8	ATTTGTATAATT	5,7
					−199	TAATCTAAAATT	4,6
clpB	**−8**	**TATGTTATAATT**	**9,5**				
				gyrA	**−8**	**TATGGTATAATT**	**11**
rpsJ	**−8**	**TTGTGTATAATT**	**8,3**		−142	TGTGGTATAATA	9,1
P97	+87	TATTATATGATT	4,5	*pyrH*	**−5**	**TAAAGTATAATA**	**6,3**
	−7	**AAAAGTATAATT**	**4,2**				
				ktrA	−191	AGTGATATAATT	6,4
glyA	**−7**	**TAGTGTATAATG**	**9,4**		**−6**	**TGGTCTACAATT**	**5,5**
	+3	ATGTGTAAAATT	5				
				rplJ	**−6**	**TCTGGTATAATT**	**10**
0225	**−7**	**TATGTTAAAATT**	**7,5**		−131	TATTATATAATA	7,8
P97-like	**−7**	**TATGGTATAATT**	**11**	*dam*	**−7**	**GCTTATATAATT**	**5,9**
	−88	TCTTGAATAATT	5,4		+35	TTTTGTAAAATG	4,3
0279	**−6**	**TAGTGTAAAATT**	**8,4**	*leuS*	**−6**	**TATGCTATAATT**	**9,5**
	−121	TAAGTTATAATA	6,7				
				P146	**−6**	**TATAGTATAATT**	**9,7**
glpK	**−6**	**AATGTTATAATT**	**6,9**		−119	TATGTTATAATT	9,5
					+21	TCTATTATAATA	6,3

Nota: nesta tabela constam apenas promotores preditos que receberam valor de peso ≥ 4,2.
† Em negrito estão as informações referentes aos promotores experimentalmente definidos para cada gene.
[a] Posição do nucleotídeo da extremidade 3' do promotor em relação ao TSS (+1). Números negativos - promotor localizado a montante do TSS. Números positivos - promotor localizado a jusante do TSS.
[b] Seqüências promotoras preditas pelo programa Matrix-Scan utilizando a PSSM de 12 colunas e a ordem 1 de Markov como modelo de *background*.
[c] Valor de peso (VP) atribuído às seqüências promotoras preditas pelo programa Matrix-Scan utilizando a PSSM de 12 colunas e a ordem 1 de Markov como modelo de *background*.
* Não foram identificadas seqüências com valor de peso > 0, a montante do gene *recA*.

A localização dos promotores em relação ao códon de iniciação também foi examinada. Esta análise revelou que 67,5% dos 201 promotores preditos estão localizados entre 1 e 100 bases a montante do códon de iniciação, estando a maioria a 25–50 bases a

montante **[FIG. 4.11]**. No entanto, 16 promotores foram encontrados dentro de 14 CDSs.

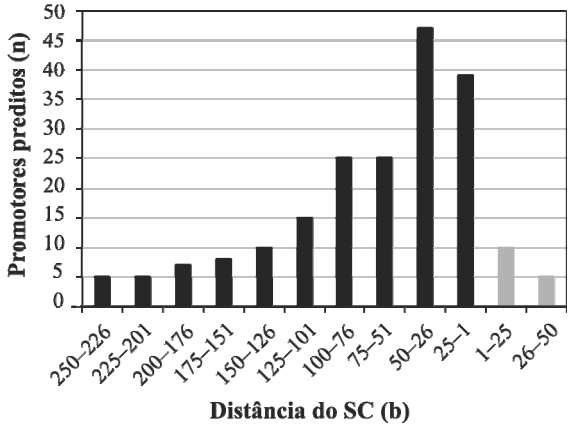
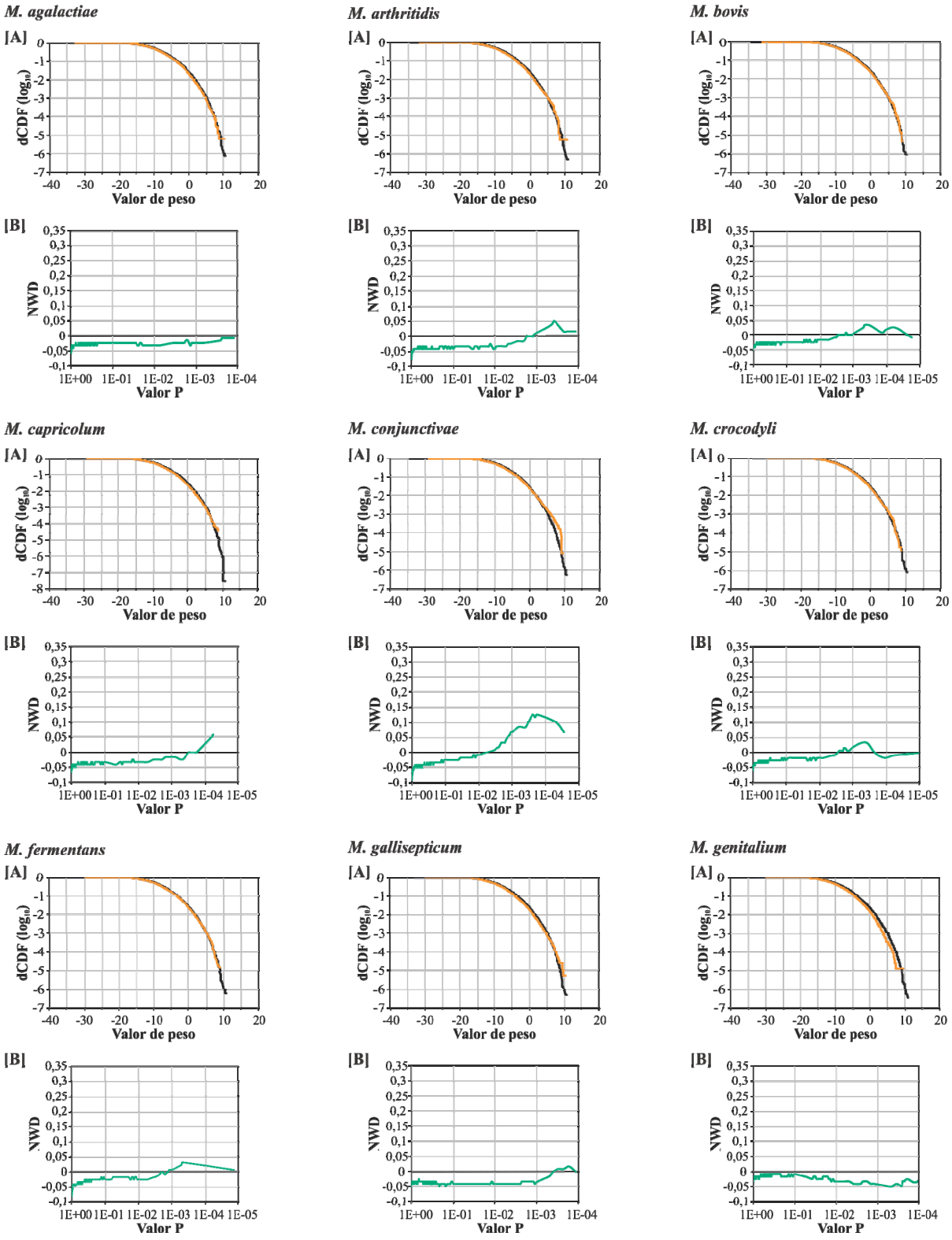


FIGURA 4.11 Localização dos promotores preditos em relação ao códon de iniciação.
Localização dos 201 promotores preditos, que apresentaram valor de peso ≥ 6,5. A distância é referente ao número de bases existentes entre o elemento −10 e o códon de iniciação (start codon, SC) dos genes anotados no genoma de *M. hyopneumoniae*. Barras pretas - bases a montante do códon de iniciação. Barras cinza - bases a jusante do códon de iniciação.

4.10 Desempenho da PSSM de 12 colunas na predição de promotores nas demais espécies de *Mycoplasma*

Embora a PSSM de 12 colunas tenha sido desenvolvida para predição de promotores em *M. hyopneumoniae* 7448, seu desempenho foi avaliado em outras 21 espécies de *Mycoplasma* e em duas outras cepas de *M. hyopneumoniae*. Nesta análise, foram calculadas as distribuições de valores de peso teórica e empírica; e a diferença entre elas foi verificada através das curvas NWD **[FIG. 4.12]**.

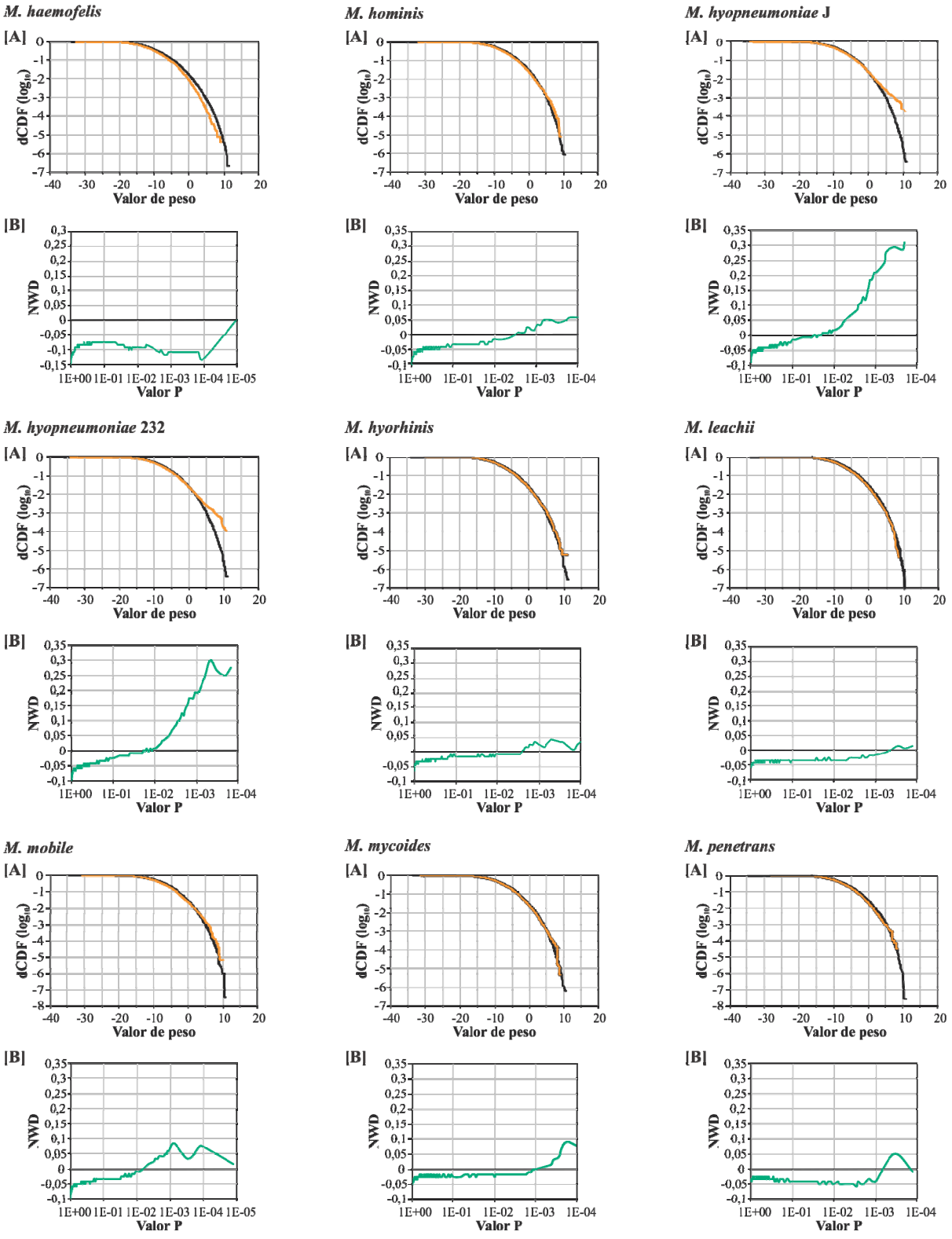
Similarmente ao realizado em *M. hyopneumoniae*, as distribuições empíricas e teóricas foram calculadas com base nas seqüências genômicas de cada organismo. As distribuições empíricas foram obtidas a partir de seqüências compostas por até 250 bases a montante e 50 bases a jusante do códon de iniciação de todas as CDSs. As distribuições teóricas foram geradas com base nas seqüências não-codificantes localizadas a montante de todos os genes.



(parte 1 de 3)

FIGURA 4.12 Desempenho da PSSM de 12 colunas na predição de promotores nas demais espécies de *Mycoplasma*.

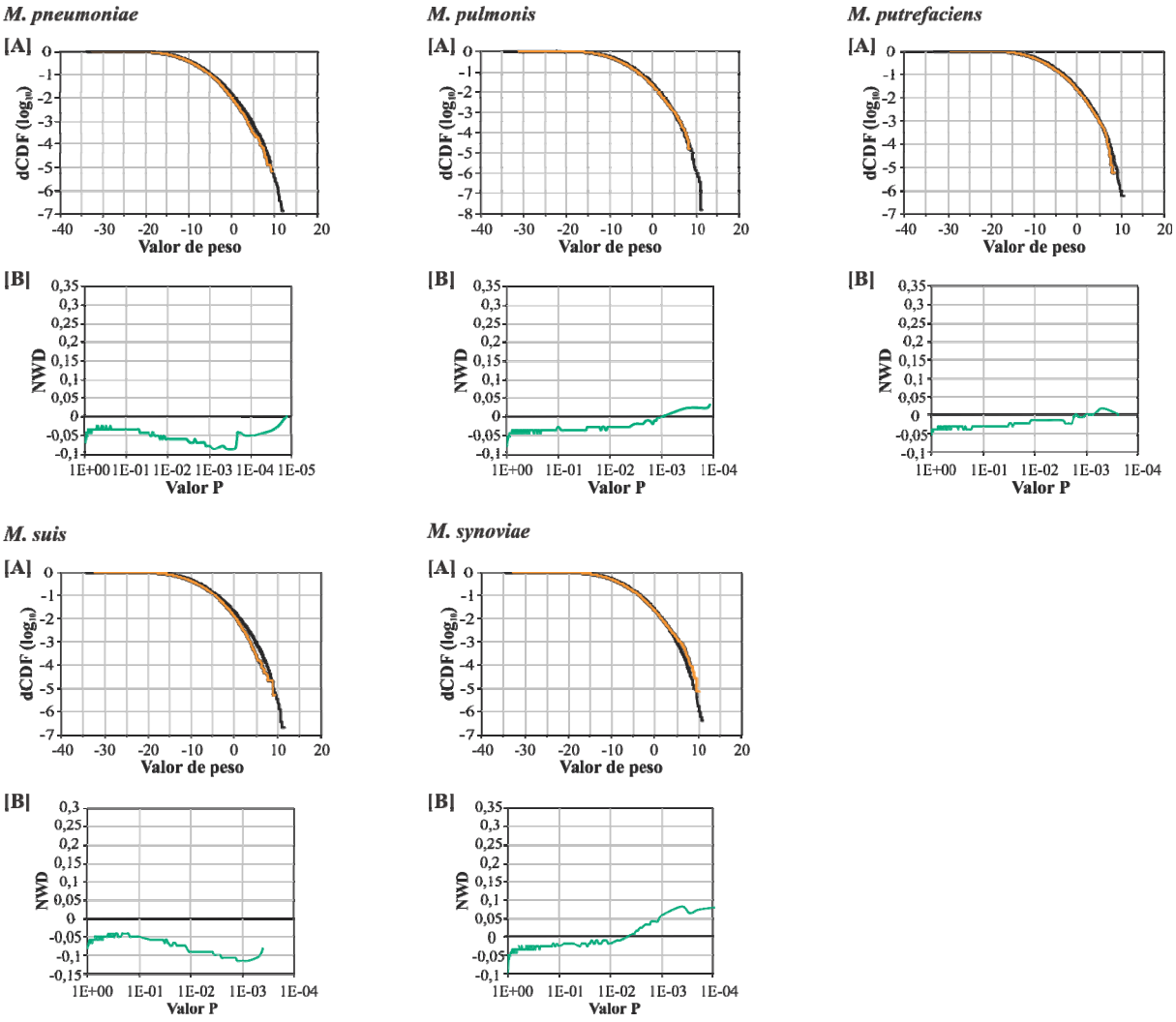
[A] Gráfico mostrando as curvas de distribuição teórica (linha preta) e empírica (linha laranja), obtidas com a PSSM de 12 colunas associada com a ordem 1 de Markov como modelo de *background*. A função de distribuição cumulativa decrescente (dCDF, ordenada) indica o valor P, ou seja, a probabilidade de observar um sítio com pontuação igual ou maior a um determinado valor de peso (abscissa). [B] Gráfico mostrando a diferença de valores de peso normalizada (NWD) calculada a partir das distribuições teórica e empírica. Quanto maior o valor de NWD, melhor é a capacidade preditiva da matriz.



(parte 2 de 3)

FIGURA 4.12 Desempenho da PSSM de 12 colunas na predição de promotores nas demais espécies de *Mycoplasma*.

[A] Gráfico mostrando as curvas de distribuição teórica (linha preta) e empírica (linha laranja), obtidas com a PSSM de 12 colunas associada com a ordem 1 de Markov como modelo de *background*. A função de distribuição cumulativa decrescente (dCDF, ordenada) indica o valor P, ou seja, a probabilidade de observar um sítio com pontuação igual ou maior a um determinado valor de peso (abscissa). [B] Gráfico mostrando a diferença de valores de peso normalizada (NWD) calculada a partir das distribuições teórica e empírica. Quanto maior o valor de NWD, melhor é a capacidade preditiva da matriz.



(parte 3 de 3)

FIGURA 4.12 Desempenho da PSSM de 12 colunas na predição de promotores nas demais espécies de *Mycoplasma*.
[A] Gráfico mostrando as curvas de distribuição teórica (linha preta) e empírica (linha laranja), obtidas com a PSSM de 12 colunas associada com a ordem 1 de Markov como modelo de *background*. A função de distribuição cumulativa decrescente (dCDF, ordenada) indica o valor P, ou seja, a probabilidade de observar um sítio com pontuação igual ou maior a um determinado valor de peso (abscissa). [B] Gráfico mostrando a diferença de valores de peso normalizada (NWD) calculada a partir das distribuições teórica e empírica. Quanto maior o valor de NWD, melhor é a capacidade preditiva da matriz.

Os gráficos da **FIG. 4.12** demonstram que, com exceção das cepas J e 232 de *M. hyopneumoniae*, em nenhuma outra espécie, a PSSM de 12 colunas apresentou desempenho semelhante ao obtido com *M. hyopneumoniae* 7448. Na maioria dos organismos valores de NWD ficaram abaixo ou um pouco acima de zero. No entanto, a matriz alcançou os maiores valores de NWD em *M. conjunctivae*, a qual, dentre as 21 espécies diferentes analisadas, é a mais próxima filogeneticamente de *M. hyopneumoniae* **[FIG. 4.12 B]**.

5. DISCUSSÃO

Com a conclusão do seqüenciamento do genoma de várias espécies de *Mycoplasma*, tem crescido o interesse pela identificação das seqüências envolvidas no início da transcrição, pois estas são fundamentais para a compreensão dos mecanismos que regulam a expressão gênica destes microrganismos. Entretanto, o estudo das regiões promotoras de micoplasmas é dificultoso, uma vez que as atuais ferramentas de bioinformática têm sido ineficientes em fazer a predição, e que são escassas as metodologias que possibilitam a caracterização experimental destas regiões.

Diferentemente de outros molicutes como *Mycoplasma pulmonis*, *Mycoplasma arthritidis*, *Mycoplasma capricolum* e *Mycoplasma pneumoniae*, *Mycoplasma hyopneumoniae* não possui um sistema-repórter estabelecido que permita o estudo de promotores *in vivo* (DYBVIG *et al*., 2000; JANIS *et al*., 2005; HALBEDEL & STULKE, 2006). Apesar das vantagens de se estudar seqüências controladoras no próprio organismo, existem alguns inconvenientes que prejudicam a aplicação desta prática nas espécies que compõe esta classe. Como demonstrado por Lartigue *et al*. (2003), os vetores replicativos de micoplasmas, geralmente, são espécie-específicos, podendo haver incompatibilidade com vetores portadores de origens de replicação heterólogas. Além disso, plasmídeos contendo *oriC* homólogas podem se integrar na *oriC* do cromossomo do hospedeiro através de recombinação homóloga (RENAUDIN *et al*., 1995).

Outra dificuldade em se trabalhar com micoplasmas deve-se ao crescimento fastidioso destas bactérias. *M. hyopneumoniae* apresenta crescimento lento, sendo que as colônias tornam-se visíveis somente após 10 dias de incubação, quando estão com aproximadamente 0,25 a 1 mm de diâmetro (ROSS, 1999). O crescimento é feito em meio

Friis, o qual é oneroso, e quando cultivado em meio sólido, os níveis de CO_2 devem ficar entre 5-10% (FRIIS, 1974).

O trabalho desenvolvido por Lopes (2007) teve como objetivos construir um vetor contendo a *oriC* de *M. hyopneumoniae* que fosse replicativo e estável, e estabelecer um protocolo de eletrotransformação para este organismo. O plasmídeo gerado, pOSTM, contendo *tetM* como gene-repórter, foi utilizado em experimentos de transformação. Como já era esperado para molicutes, foi obtida baixa freqüência de transformantes. Não foi possível, ainda, estabelecer transformantes estáveis, pois não apresentaram multiplicação após o segundo repique. A ocorrência de integração do plasmídeo no cromossomo, a instabilidade do vetor devido à presença em duplicata da região *oriC* ou a morte das células transformadas após um período prolongado de cultivo, podem ser explicações para a perda precoce dos transformantes (LOPES, 2007).

Como alternativa, Weber (2007) propôs-se a purificar o fator σ de *M. hyopneumoniae* com o intuito de, posteriormente, reconstituir a RNA polimerase holoenzima dessa bactéria, propiciando, assim, o estudo de promotores através de metodologias *in vitro*. Após a mutagênese de três códons TGA do gene *rpoD*, que possibilitaram a super-expressão heteróloga de σ de *M. hyopneumoniae* em *Escherichia coli*, a proteína foi obtida apenas na fração insolúvel, o que não permitiu sua purificação.

Tendo em vista as dificuldades encontradas no desenvolvimento de metodologias para o estudo funcional dos promotores de *M. hyopneumoniae*, a aplicação de abordagens computacionais torna-se oportuna. Contudo, tanto a caracterização quanto o reconhecimento *in silico* das seqüências regulatórias dessa espécie são limitados pela falta de promotores experimentalmente caracterizados no gênero. Visando suprir essa carência, neste trabalho, os sítios de início de transcrição de 23 genes de *M. hyopneumoniae* foram identificados. A partir

desses dados, ferramentas de bioinformática puderam ser empregadas na determinação e na predição de regiões promotoras desse patógeno.

Para identificar os sítios de início da transcrição (TSSs), optou-se pela utilização da técnica de 5' RLM-RACE (BENSING *et al*., 1996). Embora *primer extension* seja a metodologia comumente utilizada para esse fim, 5' RLM-RACE possui vantagens significativas. Dentre elas destacam-se, principalmente, a detecção das extremidades 5' de RNAs com baixo nível de expressão e a diferenciação de transcritos íntegros e processados. Enquanto o resultado do *primer extension* é proveniente apenas de uma reação de transcrição reversa, sendo, portanto, proporcional a concentração do RNA alvo (SAMBROOK & RUSSELL, 2001), o 5' RLM-RACE, além da transcrição reversa, vale-se de etapas de amplificação, facilitando o estudo de RNAs raros (BENSING *et al*., 1996). Como conseqüência, em uma reação de *primer extension*, o montante de RNA total utilizado pode variar entre 10 – 150 μg (SAMBROOK & RUSSELL, 2001), ao passo que, no 5' RLM-RACE, esses valores não ultrapassam 10 μg (BENSING *et al*., 1996). Outra desvantagem da técnica de *primer extension* é que ela não permite distinguir entre RNAs primários e processados, o que, na técnica de 5' RLM-RACE, é feito com base no estado de fosforilação das extremidades 5' dos transcritos (BENSING *et al*., 1996).

A análise da extremidade 5' dos transcritos de *M. hyopneumoniae* mostrou que a maioria deles iniciava com uma purina. O mesmo foi observado em outras bactérias (HAWLEY & MCCLURE, 1983; HELMANN, 1995; WEINER *et al*., 2000; VOGEL *et al*., 2003), corroborando com dados que mostram que a RNAP inicia a transcrição preferencialmente em sítios onde ocorrem adeninas ou guaninas (JEONG & KANG, 1994; IMBURGIO *et al*., 2000; WALKER & OSUNA, 2002). Sendo assim, a preferência por purinas juntamente com outros fatores como as seqüências nas proximidades da região de iniciação, a localização do elemento −35, as alterações na concentração intracelular dos

nucleosídeos trifosfatados e, principalmente, a distância em relação ao elemento −10, pode contribuir para a definição do TSS (JEONG & KANG, 1994; WALKER & OSUNA, 2002).

Outra característica verificada em muitos dos genes investigados foi a variação do tamanho da extremidade 5' de seus transcritos, indicando a ocorrência de TSSs heterogêneos. A variação de tamanho entre os transcritos apresentou duas configurações diferentes: os nucleotídeos extras (i) não correspondiam ou (ii) correspondiam à seqüência genômica. Na primeira configuração, os nucleotídeos adicionais sempre correspondiam a adeninas, as quais estavam inseridas em extremidades que já continham, no mínimo, outras três adeninas. De acordo com essa observação, é provável que durante a transcrição desses genes tenha havido um evento de *transcriptional slippage* – processo no qual a RNA polimerase adiciona nucleotídeos a mais, repetidamente, na extremidade 3' do transcrito nascente, normalmente dentro de seqüências homopoliméricas (TURNBOUGH, 2011). Diferentemente, na segunda configuração, onde os nucleotídeos extras eram idênticos aos dados do genoma, a heterogeneidade pode ser indicativo da existência de TSSs alternativos.

A heterogeneidade na extremidade 5' de transcritos também foi descrita para grande parte dos genes estudados em *M. pneumoniae* (WEINER *et al*., 2000). Segundo Weiner *et al*. (2000), a variação observada nessa espécie não era conseqüência de um processo de *transcriptional slippage*, mas resultado do início da transcrição entre 1 a 4 bases a montante do principal sítio de iniciação. Embora esse tipo de heterogeneidade corresponda a uma das configurações acima citadas para *M. hyopneumoniae*, algumas diferenças podem ser apontadas entre as duas espécies. Uma delas é o fato de que todos os genes de *M. pneumoniae*, nos quais a variação da extremidade 5' foi detectada, apresentavam TSSs consecutivos em decorrência dos transcritos diferirem sempre em uma base (WEINER *et al*., 2000). O mesmo pôde ser observado em alguns dos genes de *M. hyopneumoniae*, como por exemplo, *glpK*, cujos possíveis TSSs ocorriam a 27, 26 e 25 bases a montante do códon de

iniciação. Porém, nem todos os genes de *M. hyopneumoniae* possuíam "TSSs consecutivos", como no caso de *recA*, em que os TSSs ocorriam a 72 e 67 bases a montante do códon de iniciação. Outra diferença é que em *M. pneumoniae* os transcritos mais abundantes para cada gene eram os transcritos menores (WEINER *et al*., 2000), enquanto que, em *M. hyopneumoniae*, eram os transcritos maiores.

Além de uma alta proporção de transcritos com extremidades 5' heterogêneas, *M. hyopneumoniae* e *M. pneumoniae* também compartilham a peculiaridade de possuírem vários transcritos contendo apenas alguns poucos nucleotídeos na região 5'-UTR (WEINER *et al*., 2000). mRNAs como esses são chamados de *leaderless*, visto que são quase inteiramente ou totalmente desprovidos de uma seqüência líder (MOLL *et al*., 2002). Como conseqüência, essas moléculas não possuem os sinais que contribuem para a ligação dos ribossomos. No entanto, os mRNAs *leaderless* são eficientemente traduzidos nos três domínios da vida, mas, embora sejam comumente encontrados em Archaea, ainda são considerados raros entre os eucariotos e as bactérias (GRILL *et al*., 2000; BENELLI & LONDEI, 2009). Portanto, a incidência incomum de transcritos sem uma seqüência líder em *Mycoplasma* spp. poderia ser decorrente da adaptação a um genoma mínimo, com o objetivo de reduzir o espaço genômico necessário para a iniciação da tradução.

Posteriormente à determinação dos TSSs, as seqüências promotoras foram localizadas através da utilização da ferramenta Local-Word-Analysis, a qual permite a descoberta *ab initio* de sinais funcionais em seqüências biológicas (DEFRANCE *et al*., 2008). Tipicamente aplicado para detectar elementos cis-atuantes em promotores de genes co-regulados, esse programa se vale de dois critérios principais para selecionar motivos relevantes: o de representatividade, que avalia se a freqüência com que um motivo é encontrado entre as seqüências é maior do que esperado ao acaso; e o de posicionamento, que considera a concentração do motivo em uma posição específica em relação a um ponto de

referência (DEFRANCE *et al*., 2008). Portanto, neste trabalho, os promotores de *M. hyopneumoniae* foram identificados essencialmente com base na conservação tanto da seqüência quanto da localização dos motivos reconhecidos a montante dos TSSs.

De forma diferente, Weiner *et al*. (2000) identificaram as seqüências promotoras de *M. pneumoniae*. Nesse estudo, a análise da região 5' ao TSS dos genes foi feita através da utilização de matrizes derivadas a partir de dados de promotores de *E. coli*, as quais foram ajustadas de acordo com o conteúdo de G+C genômico de *M. pneumoniae*. Essa adaptação foi sugerida por Hertz & Stormo (1996) com o intuito de reduzir o viés ocasionado pela diferença da composição nucleotídica existente entre os genomas dos organismos. Como as matrizes de *E. coli* utilizadas compreendiam as regiões promotoras −35 e −10, invariavelmente, ambas estavam presentes em todos os promotores identificados em *M. pneumoniae*, ainda que, apenas um consenso fraco tenha sido observado para o elemento −35 nessa bactéria (WEINER *et al*., 2000).

O emprego interespecífico de matrizes parte do pressuposto que diferentes espécies necessariamente compartilham os mesmos sinais promotores, o que nem sempre é verdadeiro (PETERSEN *et al*., 2003; VOGEL *et al*., 2003). Diante da tendenciosidade inerente à utilização de matrizes heterólogas, o emprego de uma metodologia *ab initio* na detecção das seqüências promotoras parece ser mais apropriado, pois contempla a diversidade existente entre os genomas bacterianos. Essa abordagem ainda é mais pertinente se ponderarmos que *Mycoplasma* spp. são bastante diferentes de *E. coli*, não se restringindo apenas à variação do conteúdo de G+C de seus genomas.

Assim como nos demais micoplasmas, somente um fator σ^{70} foi identificado no genoma de *M. hyopneumoniae*. O promotor arquétipo reconhecido pelos fatores σ^{70} é composto por dois sítios principais: o elemento −35 ($^{-35}$TTGACA^{-30}) e o elemento −10 ($^{-12}$TATAAT^{-7}). Embora ambos os elementos tenham sido identificados nos promotores de

M. pneumoniae, em *M. hyopneumoniae*, apenas os elementos −10 foram observados a montante dos TSSs. Promotores contendo o elemento −10 e que não apresentam uma região −35 conservada já foram descritos para outras espécies que possuem baixo conteúdo de G+C (WOSTEN *et al*., 1998b; PETERSEN *et al*., 2003). Tem sido sugerido que tanto a degeneração de sinais regulatórios, quanto a presença de conteúdo de G+C extremamente baixo, seriam decorrentes da redução maciça do genoma pela qual alguns organismos – comumente hospedeiro-dependentes, haveriam passado (MORAN & PLAGUE, 2004; HUERTA *et al*., 2006).

Apesar de elementos −35 não terem sido detectados, a região −10 dos promotores de *M. hyopneumoniae* mostrou-se bastante conservada, sendo, na maioria dos promotores, idêntica ou possuindo cinco dos seis nucleotídeos correspondentes à seqüência consenso. Esse resultado corrobora com a análise de Mitchell *et al*. (2003), que ao averiguarem 554 promotores de *E. coli*, observaram que vários deles pareciam compensar a ocorrência de um elemento −35 pouco conservado com um elemento −10 mais próximo ao consenso $^{-12}$TATAAT^{-7}. Estudos indicam que enquanto a região −10 parece ser essencial durante a iniciação da transcrição, o elemento −35 não é indispensável, podendo sua função ser desempenhada por proteínas ativadoras ou por outras seqüências promotoras, tais como os elementos −10 estendidos e/ou elementos *upstream* (UP) (HUERTA *et al*., 2006; HOOK-BARNARD & HINTON, 2007).

O motivo 5'-TG-3', principal determinante dos elementos −10 estendidos, estava presente em 48% dos promotores experimentalmente caracterizados de *M. hyopneumoniae*. Esta proporção é muito semelhante à encontrada em *Bacillus subtilis*, na qual aproximadamente 45% dos promotores possuem este elemento (JARMER *et al*., 2001). A análise de promotores −10 estendidos mostrou que o motivo 5'-TG-3' contribui para uma atividade promotora ótima. Foi constatado que os elementos −10 estendidos compensam a

baixa conservação dos hexanucleotídeos −10 e −35 na atividade dos promotores σ^{70} (MITCHELL *et al*., 2003), podendo até mesmo suprir a completa ausência do elemento −35 (HOOK-BARNARD & HINTON, 2007). Embora relativamente freqüentes em *M. hyopneumoniae* e também encontrados nos promotores de espécies como *E. coli* e *Campylobacter jejuni*, os elementos −10 estendidos não foram observados nos promotores de *M. pneumoniae* (WEINER *et al*., 2000; GÜELL *et al*., 2009).

A montante da região −10 dos promotores de *M. hyopneumoniae*, bem como em *C. jejuni*, foram encontradas seqüências periódicas AT-ricas que também podem ter participação no início da transcrição. Petersen *et al*. (2003) sugeriram que essas regiões poderiam atuar como sítios de ligação específicos ou influenciar na curvatura do DNA. Outra hipótese seria que estas seqüências estariam relacionadas aos elementos UP, os quais estão envolvidos com o reconhecimento e a atividade de promotores (HOOK-BARNARD & HINTON, 2007). Os elementos UP têm sido identificados em diversas espécies bacterianas e, provavelmente por serem seqüências ricas em A+T, apresentam maior incidência em organismos com baixo conteúdo de G+C (DEKHTYAR *et al*., 2008). A importância desses elementos foi investigada, mostrando que sua presença aumenta significativamente a transcrição de alguns promotores e também auxilia a transcrição em promotores em que não há um elemento −35 reconhecível (HOOK-BARNARD & HINTON, 2007). Apesar de que promotores compostos apenas pelos elementos UP e −10 ainda não tenham sido identificados, já foi verificado que essa possibilidade é funcionalmente viável (ORSINI *et al*., 2004; MIROSLAVOVA & BUSBY, 2006). Portanto, em organismos AT-ricos, tais como *M. hyopneumoniae*, é possível que as extensões AT-ricas atuem como elementos UP, o que poderia minimizar a necessidade dos hexanucleotídeos −35.

A comparação entre promotores de diferentes espécies evidenciou uma maior conservação interespecífica do elemento −10 em relação ao −35. Isso corrobora com a idéia

de que o elemento −10 é fundamental, à medida que o −35 é passível de ser substituído (HOOK-BARNARD & HINTON, 2007). A ausência do elemento −35 nos *logos* de seqüências das espécies AT-ricas reforça o sugerido por Huerta *et al*. (2006), de que esses organismos tenderiam a perder os sinais promotores. Entretanto, a carência do elemento −35 estaria sendo suprida pela ocorrência dos elementos −10 estendidos e UP, os quais estão representados nos *logos* de seqüências dos promotores de *C. jejuni* e *M. hyopneumoniae*.

Embora ambas as espécies de *Mycoplasma* tenham em comum um elemento −10 com boa conservação, os promotores de *M. hyopneumoniae* são particularmente semelhantes aos de *C. jejuni*. Além de não apresentarem um elemento −35 e de possuírem uma região −10 estendida, eles também têm extensões periódicas AT-ricas a montante da região −10. Tendo em vista que *M. hyopneumoniae* (Tenericutes) e *C. jejuni* (Proteobacteria) são filogeneticamente distantes, a semelhança de seus promotores provavelmente indica convergência evolutiva, a qual poderia ser conseqüência do comportamento patogênico e do alto conteúdo genômico de A+T (70%) desses microorganismos.

A conservação das regiões de σ^{70} envolvidas com o reconhecimento das seqüências promotoras de diferentes espécies bacterianas, aparentemente, reflete a conservação observada durante a comparação de seus elementos promotores. Enfatizando a importância do elemento −10, a região 2.4 de σ^{70}, responsável por reconhecer esse hexanucleotídeo, foi a que apresentou maior conservação. Entretanto, a menor conservação das regiões 3.0 e 4.2, as quais interagem com os elementos −10 estendidos e −35, respectivamente, condiz com o fato de que esses elementos não estão presentes no *logo* de seqüência dos promotores de todas as espécies. Considerando especificamente a região 4.2 de σ^{70}, parece haver correlação entre a taxa de substituição dos aminoácidos e a presença do elemento −35 no promotor. Em *Sinorhizobium meliloti* e *B. subtilis*, onde o elemento −35 é bem evidente, a taxa de substituição não passa de 0,11, enquanto que, nas espécies onde esse

elemento não está representado, foram obtidos os maiores índices: 0,38 em *C. jejuni*, 0,52 em *M. pneumoniae* e 0,7 em *M. hyopneumoniae*. Essa relação sugere que possa estar havendo um processo de degeneração da região 4.2 da subunidade σ^{70} dessas bactérias devido a uma baixa pressão seletiva decorrente de uma menor ocorrência de elementos promotores −35 em genomas AT-ricos.

Com a finalidade de detectar possíveis seqüências promotoras ao longo das regiões intergênicas de *M. hyopneumoniae*, três matrizes de diferentes tamanhos foram construídas com base nos 23 promotores caracterizados. A menor matriz era composta por 12 colunas, representando a conservação dos nucleotídeos em 12 posições. Já as PSSMs de 14 e 16 colunas se estendiam além dessas 12 posições, incluindo informações sobre a conservação de mais alguns nucleotídeos adjacentes (dois resíduos a mais na PSSM de 14 colunas e quatro resíduos a mais na PSSM de 16 colunas). Dentre as três matrizes, a PSSM de 12 colunas foi a que apresentou a maior capacidade preditiva, indicando que as posições adicionais abrangidas pelas PSSMs de 14 e 16 colunas não eram informativas e que não teriam um significado funcional relevante.

O desempenho de uma PSSM depende de um modelo de *background* apropriado. Basicamente, esse parâmetro expressa a freqüência de cada nucleotídeo no *background*, informação que é aplicada tanto no cálculo dos valores de peso atribuídos aos segmentos de seqüência durante a predição de promotores, como na obtenção da seqüência randômica utilizada no cálculo da curva teórica (MEDINA-RIVERA *et al*., 2011). Usualmente, os modelos probabilísticos utilizados na detecção de sítios de ligação de fatores de transcrição usavam um modelo de *background* simples, baseado na freqüência dos nucleotídeos A, C, G e T existente no conjunto de seqüências que representa as seqüências investigadas pela matriz (o *background*). Entretanto, um modelo de *background* baseado na freqüência independente de cada nucleotídeo precariamente reflete a estrutura complexa das seqüências genômicas.

Alternativamente, as seqüências de DNA têm sido descritas como cadeias de Markov, implicando que a ocorrência de cada nucleotídeo seria contexto-dependente (THIJS *et al.*, 2001) **[Fig. 5.1]**.

[A]

5′-ctTACTTTCATTTTTACATTTTTTTTTTTTTTTATGCTATAATTTATAGTTACTTTGTTTATTGTCAAGTGGAGGCGGG-3′

[B]

ordem 0

A	C	G	T
15	6	13	43

ordem 1

	A	C	G	T
A	2	2	4	7
C	2	0	1	3
G	2	2	4	4
T	9	2	4	29

ordem 2

	A	C	G	T
AA	0	0	1	1
CA	1	0	0	1
GA	0	0	1	1
TA	1	2	2	4
AC	0	0	0	2
CC	0	0	0	0
GC	0	0	1	1
TC	2	0	0	0
AG	1	0	1	2
CG	0	0	1	0
GG	1	1	1	0
TG	0	1	1	2
AT	2	0	1	4
CT	1	0	0	3
GT	0	1	1	2
TT	6	1	2	20

FIGURA 5.1 Cálculo da freqüência dos nucleotídeos utilizando o modelo de Markov.
A partir das seqüências utilizadas como *background* são calculadas as freqüências de cada nucleotídeo de acordo com a ordem de Markov empregada. [A] Seqüência representando as seqüências utilizadas como *background*. Nucleotídeo em negrito – indica o nucleotídeo que está sendo contabilizado; Nucleotídeos coloridos – identifica o contexto que está sendo considerado. [B] Matrizes de transição mostrando como são contabilizadas as ocorrências de cada nucleotídeo conforme a ordem de Markov. Essas matrizes costumam ser expressas em valores de freqüência relativa, porém, para um melhor entendimento, foram mantidos os valores de freqüência absoluta. Números coloridos – são referentes à contagem do nucleotídeo (coluna) em um contexto de nucleotídeo(s) que o antecede (linha). Note que a ordem 0 não considera o contexto, indicando apenas a freqüência independente de cada nucleotídeo.

Neste trabalho, o modelo de *background* empregado foi baseado no modelo de Markov e na composição nucleotídica das seqüências intergênicas de *M. hyopneumoniae*. O modelo de Markov é uma técnica estatística que considera que a probabilidade de um nucleotídeo ser observado em uma seqüência é dependente de um contexto limitado que o precede. Sendo assim, um modelo de Markov de ordem ‘m’ significa que a probabilidade de cada resíduo depende de ‘m’ resíduos anteriores a ele na seqüência (THIJS *et al.*, 2001;

DEFRANCE *et al*., 2008). Por exemplo, com a utilização de ordem 1 (m=1), a probabilidade de que um resíduo A seja observado variará de acordo com o primeiro nucleotídeo que o antecede, ou seja, a freqüência a ser contabilizada é referente à ocorrência dos dinucleotídeos AA, TA, CA e GA nas seqüências empregadas como *background* **[Fig. 5.1]**. A vantagem desse processo pôde ser constatada mediante a observação da performance das três PSSMs utilizando as ordens 0 e 1. O desempenho de todas as matrizes foi melhor com a ordem 1 (m=1), que é contexto-dependente, em comparação com a ordem 0, cuja probabilidade é referente apenas a freqüência de cada nucleotídeo (contexto-independente).

A predição de sítios de ligação, além da matriz e do modelo de *background*, requer a definição de um valor de corte. Tendo em vista que a cada segmento de seqüência analisado é atribuído um valor de peso, é fundamental estabelecer um critério que indique quais destes serão considerados como prováveis sítios de ligação (GAHATHAKURTA & STORMO, 2007). Embora a escolha do valor de corte esteja relacionada às taxas de sensibilidade e de falso-positivos, ainda é comumente realizada de forma arbitrária. Em alguns trabalhos, por exemplo, o valor de corte consiste no menor valor de peso obtido para sítios conhecidos (utilizados na construção da matriz) (VOGEL *et al*., 2003; KOHL *et al*., 2008; TRUNK *et al*., 2010). Entretanto, apesar dessa estratégia garantir uma sensibilidade de 100%, a predição de falso-positivos tende a ser muito alta. Diferentemente, outros trabalhos determinaram o valor de corte como sendo igual a dois desvios-padrão abaixo da média dos valores de peso de todos os sítios conhecidos (ROBISON *et al*., 1998). Com um valor de peso mais elevado, a ocorrência de falso-positivos é reduzida, porém muitos sítios de ligação verdadeiros acabam sendo ignorados, diminuindo a sensibilidade.

Para determinar o valor de corte para a predição de promotores de *M. hyopneumoniae* com a PSSM de 12 colunas, foi adotada uma abordagem diferente, que levava em conta a ocorrência de falso-positivos. Essa metodologia assume que promotores preditos

nas fitas anti-senso de seqüências intergênicas corresponderiam a falso-positivos e que essa proporção seria equivalente dentre os promotores preditos na fita senso (seqüências a montante dos genes, com mesma orientação, onde se espera encontrar os sítios biologicamente funcionais) (CASES *et al*., 2003). Deste modo, o valor de peso notoriamente associado a uma menor proporção de falso-positivos foi escolhido como valor de corte. Ainda, a correlação entre sensibilidade e FPR atrelada a esse valor foi considerada nessa escolha.

A busca por promotores putativos nas regiões intergênicas do genoma de *M. hyopneumoniae* mostrou que a incidência do padrão detectado pela matriz é mais freqüente do que o esperado, indicando que sua ocorrência não é ao acaso, e que, provavelmente, são capazes de promover o início da transcrição gênica. Estudos recentes, baseados em dados para os promotores σ^{70} de *E. coli*, não foram capazes de detectar estes padrões em genomas de *Mycoplasma* spp. (HUERTA *et al*., 2006; SINOQUET *et al*., 2008), e até mesmo sugeriram que a existência de promotores nestas bactérias seria questionável (SINOQUET *et al*., 2008). No entanto, como demonstrado por Weiner *et al*. (2000), a identificação de promotores em *Mycoplasma* spp. utilizando uma matriz de *E. coli* foi pouco eficiente. Portanto, o diferencial do presente estudo deve-se ao emprego de uma PSSM espécie-específica que considerou a variabilidade existente entre as espécies, evitando vieses provenientes da utilização de PSSMs heterólogas.

Através da utilização da PSSM de 12 colunas e de um valor de corte de 6,5, foram identificados promotores putativos a montante de aproximadamente 26% das CDSs de *M. hyopneumoniae*. No entanto, essa cobertura deve ser maior do que estimado *a priori*, uma vez que muitas das regiões intergênicas analisadas eram demasiadamente curtas para conter uma seqüência promotora de 12 nucleotídeos separada por quatro nucleotídeos do TSS. Além disso, Adams *et al*. (2005) sugeriu que o tamanho máximo das regiões intergênicas de um

operon de *M. hyopneumoniae* seria de aproximadamente 50 bases. Ainda, estudos mostraram que genes organizados *in tandem* com distâncias intergênicas muito maiores do que 50 bases também podem compor grandes unidades transcricionais (GARDNER & MINION, 2010; SIQUEIRA *et al*., 2011). Esses achados indicam que muitas CDSs dessa bactéria são reguladas por um promotor em comum, e que, portanto, nem todas as regiões codificantes, necessariamente, devam estar acompanhadas por um promotor imediatamente localizado na região 5' que a precede.

A organização transcricional de *M. hyopneumoniae* também explicaria o fato de que seqüências promotoras foram identificadas em apenas 93 das 515 CDSs (18%) cujo gene a montante apresenta mesma orientação. Isso também seria resultado da distribuição desses genes dentro de unidades transcricionais, as quais podem ser controladas por um único promotor localizado a montante do primeiro gene. No entanto, nas CDSs em que o gene a montante tinha orientação divergente, foi possível detectar promotores em 54% delas. Essa proporção relativamente maior era esperada, uma vez que CDSs com este contexto gênico, obrigatoriamente, possuem uma região regulatória própria.

A alta freqüência de transcritos alternativos, bem como, a subdivisão das unidades transcricionais de *M. pneumoniae* em UTs menores, provavelmente, também sejam observadas em *M. hyopneumoniae*. A predição de promotores imediatamente a montante de vários genes considerados internos em UTs reforça essa possibilidade. Um exemplo é a unidade transcricional experimentalmente definida composta pelos genes *deoC*, *upp*, MHP7448_0525, *lon* e *tuf* (SIQUEIRA *et al*., 2011). Todos os genes, exceto MHP7448_0525, contêm seqüências promotoras nas suas regiões a montante (com pontuações variando de 8,4-11) (dados não mostrados). Isso corrobora os resultados de Gardner *et al*. (2010), que indicam a possibilidade de início de transcrição independente de genes que compõem uma mesma UT.

A maioria das CDSs (84%) apresentou uma seqüência promotora única, mas CDSs contendo múltiplas seqüências promotoras também foram verificadas. O gene *tuf*, por exemplo, que é conhecido por ser altamente expresso, possui três promotores na sua região a montante, dois dos quais sobrepostos (dados não mostrados). A sobreposição de sinais pode estar envolvida na promoção da transcrição. Isso pode ocorrer através do recrutamento da RNA polimerase para a seqüência promotora primária (REZNIKOFF *et al*., 1987), ou, quando esta se encontra ausente, os sítios podem corresponder a promotores fracos não-competitivos, promovendo a transcrição basal. No entanto, promotores sobrepostos também podem regular negativamente a transcrição, competindo pelas RNA polimerases (GOODRICH & MCCLURE, 1991), ou induzindo uma pausa nas etapas iniciais do alongamento durante a síntese de RNA (BRODOLIN *et al*., 2004; NICKELS *et al*., 2004).

A maioria dos promotores putativos foi encontrada entre 1 e 100 bases a montante do códon de iniciação. Este resultado é congruente com o observado em muitos estudos realizados em diferentes espécies bacterianas (PETERSEN *et al*., 2003). Algumas seqüências promotoras preditas foram encontradas dentro das CDSs. Isso pode ser atribuído à anotação imprecisa dos códons de iniciação desses genes, ou a sinais putativos intragênicos com função regulatória desconhecida.

Embora, neste estudo, uma busca abrangente por promotores tenha sido realizada, muitos promotores putativos não foram identificados devido aos critérios utilizados para a predição. Nesse contexto, a principal limitação foi a pontuação limiar empregada ($\geq$ 6,5). Aproximadamente 30% dos promotores definidos experimentalmente, neste estudo, não foram detectados utilizando este valor de corte. A menor pontuação obtida para estes promotores foi de 4,2, no entanto, neste limiar, cerca de metade das seqüências identificadas nas regiões intergênicas de *M. hyopneumoniae* foram estimadas como falso-positivas.

A grande quantidade de seqüências que apresentam valor de peso ≥ 4,2 (3,46 sítios a cada 1.000 bases) deve ser considerada no questionamento de como RNA polimerase faria para distinguir entre os promotores verdadeiros e os falso-positivos. Como *M. hyopneumoniae* tem apenas um pequeno número de proteínas reguladoras conhecidas (MINION *et al*., 2004), especula-se que a maior parte das seqüências que tem pontuação ≥ 4,2 seria promotor verdadeiro. Portanto, duas principais conjecturas poderiam ser consideradas: o contexto onde estas seqüências estão imersas seria determinante para uma iniciação da transcrição apropriada e/ou talvez haja transcrição generalizada em *M. hyopneumoniae*. De acordo, apesar de estudos mostrarem que *M. hyopneumoniae* é capaz de controlar a transcrição (WEINER, *et al*., 2003; MADSEN *et al*., 2006a; MADSEN *et al*., 2006b; SCHAFER *et al*., 2007; ONEAL *et al*., 2008), Gardner *et al*. (2010) demonstraram que a maioria das regiões intergênicas desse organismo são intensamente transcritas.

A possibilidade da PSSM de 12 colunas ser utilizada na busca por promotores em nas demais espécies de *Mycoplasma* foi avaliada. Por pertencerem a um mesmo gênero e possuírem apenas um único fator σ, presumiu-se que seus promotores compartilhariam as mesmas características. Entretanto, somente a espécie *M. conjunctivae*, mais próxima filogeneticamente de *M. hyopneumoniae*, apresentou freqüência de promotores putativos maior do que a esperada ao acaso. Esse resultado, provavelmente, deve-se às diferenças na composição nucleotídica dos promotores, intrínseca de cada espécie. Isso reforça a necessidade de se utilizar matrizes espécie-específicas, pois como mencionado anteriormente, PSSMs heterólogas não consideram a variabilidade interespecífica existente.

Este estudo contribuiu para a compreensão dos mecanismos de transcrição em *M. hyopneumoniae*, uma vez que identificou os elementos básicos envolvidos no início da transcrição e verificou a sua distribuição nas regiões a montante dos genes codificadores de proteínas desta espécie. Características aparentemente raras em outras espécies bacterianas,

incluindo TSSs heterogêneos e a falta de uma região líder não-traduzida que contivesse um RBS, parecem ser comuns em *M. hyopneumoniae*. A caracterização dos promotores evidenciou a ocorrência de elementos −10 estendidos e seqüências periódicas AT-ricas, não descritas antes para o gênero. Finalmente, a predição de promotores a montante de genes *in tandem* indica que as UTs preditas por Siqueira *et al*. (2011) devam ser subdivididas em UTs menores.

6. PERSPECTIVAS

- Correlacionar os resultados da predição, realizada neste trabalho, com análises de *footprinting* filogenético para confirmar os promotores preditos;
- Desenvolver metodologias que permitam o estudo *in vivo* ou *in vitro* de promotores de *Mycoplasma hyopneumoniae*;
- Realizar estudos funcionais com os promotores preditos, visando verificar a necessidade de elementos −35 e a participação das seqüências periódicas AT-ricas no início da transcrição de *M. hyopneumoniae*;
- Correlacionar dados de intensidade de expressão de genes com a composição nucleotídica de seus promotores;
- Utilizar a PSSM, definida neste trabalho, para refinar as anotações do genoma de *M. hyopneumoniae* e para investigar promotores em *Mycoplasma flocculare*, espécie estreitamente relacionada à *M. hyopneumoniae*.

7. REFERÊNCIAS

ADAMS, C.; PITZER, J. & MINION, F. C. *In vivo* expression analysis of the P97 and P102 paralog families of *Mycoplasma hyopneumoniae*. Infect. Immun., 73(11): 7784-7787, 2005.

BAILEY, T. L. & ELKAN, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proc. Int. Conf. Intell. Syst. Mol. Biol., 228-36, 1994.

BALLEZA, E.; LOPEZ-BOJORQUEZ, L. N.; MARTINEZ-ANTONIO, A.; RESENDIS-ANTONIO, O.; LOZADA-CHAVEZ, I.; BALDERAS-MARTINEZ, Y. I.; ENCARNACION, S. & COLLADO-VIDES, J. Regulation by transcription factors in bacteria: beyond description. FEMS Microbiol. Rev., 33(1): 133-151, 2009.

BAR-NAHUM, G. & NUDLER, E. Isolation and characterization of sigma(70)-retaining transcription elongation complexes from *Escherichia coli*. Cell, 106(4): 443-451, 2001.

BASEMAN, J. B.; LANGE, M.; CRISCIMAGNA, N. L.; GIRON, J. A. & THOMAS, C. A. Interplay between mycoplasmas and host target cells. Microb. Pathog., 19(2): 105-116, 1995.

BASEMAN, J. B. & TULLY, J. G. Mycoplasmas: sophisticated, reemerging, and burdened by their notoriety. Emerg. Infect. Dis., 3(1): 21-32, 1997.

BENELLI, D. & LONDEI, P. Begin at the beginning: evolution of translational initiation. Res. Microbiol, 160(7): 493-501, 2009.

BENSING, B. A.; MEYER, B. J. & DUNNY, G. M. 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. U.S.A, 93(15): 7794-7799, 1996.

BENTLEY, S. D.; CHATER, K. F.; CERDENO-TARRAGA, A. M.; CHALLIS, G. L.; THOMSON, N. R.; JAMES, K. D.; HARRIS, D. E.; QUAIL, M. A.; KIESER, H.; HARPER, D.; BATEMAN, A.; BROWN, S.; CHANDRA, G.; CHEN, C. W.; COLLINS, M.; CRONIN, A.; FRASER, A.; GOBLE, A.; HIDALGO, J.; HORNSBY, T.; HOWARTH, S.; HUANG, C. H.; KIESER, T.; LARKE, L.; MURPHY, L.; OLIVER, K.; O'NEIL, S.; RABBINOWITSCH, E.; RAJANDREAM, M. A.; RUTHERFORD, K.; RUTTER, S.; SEEGER, K.; SAUNDERS, D.; SHARP, S.; SQUARES, R.; SQUARES, S.; TAYLOR, K.; WARREN, T.; WIETZORREK, A.; WOODWARD, J.; BARRELL, B. G.; PARKHILL, J. & HOPWOOD, D. A. Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). Nature, 417(6885): 141-147, 2002.

BHUGRA, B.; VOELKER, L. L.; ZOU, N.; YU, H. & DYBVIG, K. Mechanism of antigenic variation in *Mycoplasma pulmonis*: interwoven, site-specific DNA inversions. Mol. Microbiol., 18(4): 703-714, 1995.

BLAND, C.; NEWSOME, A. S. & MARKOVETS, A. A. Promoter prediction in *E. coli* based on SIDD profiles and Artificial Neural Networks. BMC Bioinformatics., 11 Suppl 6S17-2010.

BLATTNER, F. R.; PLUNKETT, G., III; BLOCH, C. A.; PERNA, N. T.; BURLAND, V.; RILEY, M.; COLLADO-VIDES, J.; GLASNER, J. D.; RODE, C. K.; MAYHEW, G. F.; GREGOR, J.; DAVIS, N. W.; KIRKPATRICK, H. A.; GOEDEN, M. A.; ROSE, D. J.; MAU, B. & SHAO, Y. The complete genome sequence of *Escherichia coli* K-12. Science, 277(5331): 1453-1474, 1997.

BORUKHOV, S. & SEVERINOV, K. Role of the RNA polymerase sigma subunit in transcription initiation. Res. Microbiol., 153(9): 557-562, 2002.

BORUKHOV, S. & NUDLER, E. RNA polymerase holoenzyme: structure, function and biological implications. Curr. Opin. Microbiol, 6(2): 93-100, 2003.

BOVE, J. M. Molecular features of mollicutes. Clin. Infect. Dis., 17 Suppl 1S10-S31, 1993.

BRODOLIN, K.; ZENKIN, N.; MUSTAEV, A.; MAMAEVA, D. & HEUMANN, H. The sigma 70 subunit of RNA polymerase induces lacUV5 promoter-proximal pausing of transcription. Nat. Struct. Mol. Biol., 11(6): 551-557, 2004.

BROWNING, D. F. & BUSBY, S. J. The regulation of bacterial transcription initiation. Nat. Rev. Microbiol., 2(1): 57-65, 2004.

CASES, I.; USSERY, D. W. & DE, L., V The sigma54 regulon (sigmulon) of *Pseudomonas putida*. Environ. Microbiol., 5(12): 1281-1293, 2003.

CIPRIAN, A.; PIJOAN, C.; CRUZ, T.; CAMACHO, J.; TORTORA, J.; COLMENARES, G.; LOPEZ-REVILLA, R. & DE LA, G. M. *Mycoplasma hyopneumoniae* increases the susceptibility of pigs to experimental *Pasteurella multocida* pneumonia. Can. J. Vet. Res., 52(4): 434-438, 1988.

CROOKS, G. E.; HON, G.; CHANDONIA, J. M. & BRENNER, S. E. WebLogo: a sequence logo generator. Genome Res., 14(6): 1188-1190, 2004.

DARST, S. A. Bacterial RNA polymerase. Curr. Opin. Struct. Biol., 11(2): 155-162, 2001.

DEBEY, M. C. & ROSS, R. F. Ciliostasis and loss of cilia induced by *Mycoplasma hyopneumoniae* in porcine tracheal organ cultures. Infect. Immun., 62(12): 5312-5318, 1994.

DEFRANCE, M.; JANKY, R.; SAND, O. & VAN, H. J. Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. Nat. Protoc., 3(10): 1589-1603, 2008.

DEKHTYAR, M.; MORIN, A. & SAKANYAN, V. Triad pattern algorithm for predicting strong promoter candidates in bacterial genomes. BMC Bioinformatics., 9233-2008.

DHANDAYUTHAPANI, S.; RASMUSSEN, W. G. & BASEMAN, J. B. Identification of mycoplasmal promoters in *Escherichia coli* using a promoter probe vector with Green Fluorescent Protein as reporter system. Gene, 215(1): 213-222, 1998.

DJORDJEVIC, S. P.; CORDWELL, S. J.; DJORDJEVIC, M. A.; WILTON, J. & MINION, F. C. Proteolytic processing of the *Mycoplasma hyopneumoniae* cilium adhesin. Infect. Immun., 72(5): 2791-2802, 2004.

DYBVIG, K. & VOELKER, L. L. Molecular biology of mycoplasmas. Annu. Rev. Microbiol., 5025-57, 1996.

DYBVIG, K.; FRENCH, C. T. & VOELKER, L. L. Construction and use of derivatives of transposon Tn4001 that function in *Mycoplasma pulmonis* and *Mycoplasma arthritidis*. J. Bacteriol., 182(15): 4343-4347, 2000.

EDGAR, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res., 32(5): 1792-1797, 2004.

ERMOLAEVA, M. D.; WHITE, O. & SALZBERG, S. L. Prediction of operons in microbial genomes. Nucleic Acids Res., 29(5): 1216-1221, 2001.

ESTREM, S. T.; ROSS, W.; GAAL, T.; CHEN, Z. W.; NIU, W.; EBRIGHT, R. H. & GOURSE, R. L. Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. Genes Dev., 13(16): 2134-2147, 1999.

EWING, B.; HILLIER, L.; WENDL, M. C. & GREEN, P. Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. Genome Res., 8(3): 175-185, 1998.

FARNHAM, P. J. & PLATT, T. Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription *in vitro*. Nucleic Acids Res., 9(3): 563-577, 1981.

FINCH, L. R. & MITCHELL, A. Sources of nucleotides In: Maniloff, J.; McElhaney, R. N.; Finch, L. R. & Baseman, J. B. editors. Mycoplasmas: Molecular Biology and Pathogenesis. Washington, DC: Am. Soc. Microbiol., 1992. p. 211-230.

FRASER, C. M.; GOCAYNE, J. D.; WHITE, O.; ADAMS, M. D.; CLAYTON, R. A.; FLEISCHMANN, R. D.; BULT, C. J.; KERLAVAGE, A. R.; SUTTON, G.; KELLEY, J. M.; FRITCHMAN, R. D.; WEIDMAN, J. F.; SMALL, K. V.; SANDUSKY, M.; FUHRMANN, J.; NGUYEN, D.; UTTERBACK, T. R.; SAUDEK, D. M.; PHILLIPS, C. A.; MERRICK, J. M.; TOMB, J. F.; DOUGHERTY, B. A.; BOTT, K. F.; HU, P. C.; LUCIER, T. S.; PETERSON, S. N.; SMITH, H. O.; HUTCHISON, C. A., III & VENTER, J. C. The minimal gene complement of *Mycoplasma genitalium*. Science, 270(5235): 397-403, 1995.

FRIIS, N. F. Micoplasmas in pigs, with a special regard to the respiratory tract., Univ. Royal Vet. and Agricult., Copenhagen, 1974.

FRIIS, N.F. Some recommendations concerning primary isolation of *Mycoplasma suipneumoniae* and *Mycoplasma flocculare* a survey. Nord. Vet. Med., 27(6): 337-339, 1975.

GAFNY, R.; HYMAN, H. C.; RAZIN, S. & GLASER, G. Promoters of *Mycoplasma capricolum* ribosomal RNA operons: identical activities but different regulation in homologous and heterologous cells. Nucleic Acids Res., 16(1): 61-76, 1988.

GAHATHAKURTA, D. & STORMO, G. D. Finding regulatory elements in DNA sequence In: Dear, P., editors. Bioinformatics: Methods Express. Oxfordshire, UK: Scion Publisher Ltd., 2007. p. 117-139.

GARDNER, S. W. & MINION, F. C. Detection and quantification of intergenic transcription in *Mycoplasma hyopneumoniae*. Microbiology, 156(Pt 8): 2305-2315, 2010.

GHOSH, T.; BOSE, D. & ZHANG, X. Mechanisms for activating bacterial RNA polymerase. FEMS Microbiol Rev., 34(5): 611-627, 2010.

GIL, R.; SILVA, F. J.; PERETO, J. & MOYA, A. Determination of the core of a minimal bacterial gene set. Microbiol. Mol. Biol. Rev., 68(3): 518-537, 2004.

GLEW, M. D.; BASEGGIO, N.; MARKHAM, P. F.; BROWNING, G. F. & WALKER, I. D. Expression of the pMGA genes of *Mycoplasma gallisepticum* is controlled by variation in the GAA trinucleotide repeat lengths within the 5' noncoding regions. Infect. Immun., 66(12): 5833-5841, 1998.

GOODRICH, J. A. & MCCLURE, W. R. Competing promoters in prokaryotic transcription. Trends Biochem. Sci., 16(11): 394-397, 1991.

GRECH, B.; MAETSCHKE, S.; MATHEWS, S., & TIMMS, P. Genome-wide analysis of chlamydiae for promoters that phylogenetically footprint. Res. Microbiol., 158(8-9): 685-693, 2007.

GRILL, S.; GUALERZI, C. O.; LONDEI, P. & BLASI, U. Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation. EMBO J, 19(15): 4101-4110, 2000.

GRUBER, T. M. & GROSS, C. A. Multiple sigma subunits and the partitioning of bacterial transcription space. Annu. Rev. Microbiol., 57441-466, 2003.

GÜELL, M.; VAN, N., V; YUS, E.; CHEN, W. H.; LEIGH-BELL, J.; MICHALODIMITRAKIS, K.; YAMADA, T.; ARUMUGAM, M.; DOERKS, T.; KUHNER, S.; RODE, M.; SUYAMA, M.; SCHMIDT, S.; GAVIN, A. C.; BORK, P. & SERRANO, L. Transcriptome complexity in a genome-reduced bacterium. Science, 326(5957): 1268-1271, 2009.

HALBEDEL, S. & STULKE, J. Probing *in vivo* promoter activities in *Mycoplasma pneumoniae*: a system for generation of single-copy reporter constructs. Appl. Environ. Microbiol., 72(2): 1696-1699, 2006.

HALBEDEL, S.; EILERS, H.; JONAS, B.; BUSSE, J.; HECKER, M.; ENGELMANN, S. & STULKE, J. Transcription in *Mycoplasma pneumoniae*: Analysis of the Promoters of the *ackA* and *ldh* Genes. J. Mol. Biol., 2007.

HALL, T. A. BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symposium Series, 41(1): 95-98, 1999.

HAWLEY, D. K. & MCCLURE, W. R. Compilation and analysis of *Escherichia coli* promoter DNA sequences. Nucleic Acids Res., 11(8): 2237-2255, 1983.

HELMANN, J. D. 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(13): 2351-2360, 1995.

HERTZ, G. Z. & STORMO, G. D. *Escherichia coli* promoter sequences: analysis and prediction. Methods Enzymol., 27330-42, 1996.

HERTZ, G. Z. & STORMO, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics., 15(7-8): 563-577, 1999.

HOOK-BARNARD, I. G. & HINTON, D. M. Transcription initiation by mix and match elements: flexibility for polymerase binding to bacterial promoters. Gene Regul. Syst. Bio., 1275-293, 2007.

HOON, M. J.; MAKITA, Y.; NAKAI, K. & MIYANO, S. Prediction of transcriptional terminators in *Bacillus subtilis* and related species. PLoS. Comput. Biol., 1(3): e25-2005.

HUERTA, A. M. & COLLADO-VIDES, J. Sigma70 promoters in *Escherichia coli*: specific transcription in dense regions of overlapping promoter-like signals. J. Mol. Biol., 333(2): 261-278, 2003.

HUERTA, A. M.; FRANCINO, M. P.; MORETT, E., & COLLADO-VIDES, J. Selection for unequal densities of sigma 70 promoter-like signals in different regions of large bacterial genomes. PLoS. Genet., 2(11): e185-2006.

HYMAN, H. C.; GAFNY, R.; GLASER, G. & RAZIN, S. Promoter of the *Mycoplasma pneumoniae* rRNA operon. J. Bacteriol., 170(7): 3262-3268, 1988.

IMBURGIO, D.; RONG, M.; MA, K. & MCALLISTER, W. T. Studies of promoter recognition and start site selection by T7 RNA polymerase using a comprehensive collection of promoter variants. Biochemistry, 39(34): 10419-10430, 2000.

JANIS, C.; LARTIGUE, C.; FREY, J.; WROBLEWSKI, H.; THIAUCOURT, F.; BLANCHARD, A. & SIRAND-PUGNET, P. Versatile use of oriC plasmids for functional genomics of *Mycoplasma capricolum* subsp. capricolum. Appl. Environ. Microbiol., 71(6): 2888-2893, 2005.

JARMER, H.; LARSEN, T. S.; KROGH, A.; SAXILD, H. H.; BRUNAK, S. & KNUDSEN, S. Sigma A recognition sites in the *Bacillus subtilis* genome. Microbiology, 147(Pt 9): 2417-2424, 2001.

JEONG, W. & KANG, C. Start site selection at lacUV5 promoter affected by the sequence context around the initiation sites. Nucleic Acids Res., 22(22): 4667-4672, 1994.

JONES, D. T.; TAYLOR, W. R. & THORNTON, J. M. The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci., 8(3): 275-282, 1992.

KANHERE, A. & BANSAL, M. A novel method for prokaryotic promoter prediction based on DNA stability. BMC Bioinformatics., 61-2005.

KIM, K. B. & SIM, J. S. Computational detection of prokaryotic core promoters in genomic sequences. J.Microbiol., 43(5): 411-416, 2005.

KLEIN, C. C. Caracterização de genes envolvidos no metabolismo de mio-Inositol em *Mycoplasma hyopneumoniae*. Monografia (Título de Bacharel em Ciências Biológicas) - Faculdade de Ciências Biológicas, UFRGS, Porto Alegre, 2008.

KNUDTSON, K. L. & MINION, F. C. Use of lac gene fusions in the analysis of *Acholeplasma* upstream gene regulatory sequences. J. Bacteriol., 176(9): 2763-2766, 1994.

KOHL, T. A.; BAUMBACH, J.; JUNGWIRTH, B.; PUHLER, A. & TAUCH, A. The GlxR regulon of the amino acid producer *Corynebacterium glutamicum*: in *silico* and *in vitro* detection of DNA binding sites of a global transcription regulator. J. Biotechnol., 135(4): 340-350, 2008.

KUNST, F.; OGASAWARA, N.; MOSZER, I.; ALBERTINI, A. M.; ALLONI, G.; AZEVEDO, V.; BERTERO, M. G.; BESSIERES, P.; BOLOTIN, A.; BORCHERT, S.; BORRISS, R.; BOURSIER, L.; BRANS, A.; BRAUN, M.; BRIGNELL, S. C.; BRON, S.; BROUILLET, S.; BRUSCHI, C. V.; CALDWELL, B.; CAPUANO, V.; CARTER, N. M.; CHOI, S. K.; CODANI, J. J.; CONNERTON, I. F.; DANCHIN, A. & . The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. Nature, 390(6657): 249-256, 1997.

LANE, D.; PRENTKI, P. & CHANDLER, M. Use of gel retardation to analyze protein-nucleic acid interactions. Microbiol. Rev., 56(4): 509-528, 1992.

LARTIGUE, C.; BLANCHARD, A.; RENAUDIN, J.; THIAUCOURT, F. & SIRAND-PUGNET, P. Host specificity of mollicutes oriC plasmids: functional analysis of replication origin. Nucleic Acids Res., 31(22): 6610-6618, 2003.

LISSER, S. & MARGALIT, H. Compilation of *E. coli* mRNA promoter sequences. Nucleic Acids Res., 21(7): 1507-1516, 1993.

LIU, W.; FENG, Z.; FANG, L.; ZHOU, Z.; LI, Q.; LI, S.; LUO, R.; WANG, L.; CHEN, H.; SHAO, G. & XIAO, S. Complete genome sequence of *Mycoplasma hyopneumoniae* strain 168. J. Bacteriol., 193(4): 1016-1017, 2011.

LLUCH-SENAR, M.; VALLMITJANA, M.; QUEROL, E. & PINOL, J. A new promoterless reporter vector reveals antisense transcription in *Mycoplasma genitalium*. Microbiology, 153(Pt 8): 2743-2752, 2007.

LO, S. C.; HAYES, M. M.; KOTANI, H.; PIERCE, P. F.; WEAR, D. J.; NEWTON, P. B., III; TULLY, J. G. & SHIH, J. W. Adhesion onto and invasion into mammalian cells by *Mycoplasma penetrans*: a newly isolated mycoplasma from patients with AIDS. Mod. Pathol., 6(3): 276-280, 1993.

LONETTO, M.; GRIBSKOV, M. & GROSS, C. A. The sigma 70 family: sequence conservation and evolutionary relationships. J. Bacteriol., 174(12): 3843-3849, 1992.

LOPES, B. M. T. Construção de vetor *oriC* de *Mycoplasma hyopneumoniae* - uma ferramenta para estudos genéticos do agente da Pneumonia Enzoótica Suína. Dissertação (Mestrado em Ciências Veterinárias) - Programa de Pós-graduação em Ciências Veterinárias, UFRGS, Porto Alegre, 2007.

LPSN - List of Prokaryotic names with Standing in Nomenclature. Genus *Mycoplasma*. Disponível em: <http://www.bacterio.cict.fr/m/mycoplasma.html>. Acesso em: 19/01/2012

MACLELLAN, S. R.; MACLEAN, A. M., & FINAN, T. M. Promoter prediction in the rhizobia. Microbiology, 152(Pt 6): 1751-1763, 2006.

MADSEN, M. L.; NETTLETON, D.; THACKER, E. L.; EDWARDS, R. & MINION, F. C. Transcriptional profiling of *Mycoplasma hyopneumoniae* during heat shock using microarrays. Infect. Immun., 74(1): 160-166, 2006a.

MADSEN, M. L.; NETTLETON, D.; THACKER, E. L. & MINION, F. C. Transcriptional profiling of *Mycoplasma hyopneumoniae* during iron depletion using microarrays. Microbiology, 152(Pt 4): 937-944, 2006b.

MANOLUKAS, J. T.; BARILE, M. F.; CHANDLER, D. K. & POLLACK, J. D. Presence of anaplerotic reactions and transamination, and the absence of the tricarboxylic acid cycle in mollicutes. J. Gen. Microbiol., 134(3): 791-800, 1988.

MEDINA-RIVERA, A.; BREU-GOODGER, C.; THOMAS-CHOLLIER, M.; SALGADO, H.; COLLADO-VIDES, J. & VAN, H. J. Theoretical and empirical quality assessment of transcription factor-binding motifs. Nucleic Acids Res., 39(3): 808-824, 2011.

MILLER, J. H. Experiments in molecular genetics. Cold Spring Harbor, N.Y.: Cold Spring Harbor Laboratory, 1972.

MINION, F. C.; LEFKOWITZ, E. J.; MADSEN, M. L.; CLEARY, B. J.; SWARTZELL, S. M. & MAHAIRAS, G. G. The genome sequence of *Mycoplasma hyopneumoniae* strain 232, the agent of swine mycoplasmosis. J. Bacteriol., 186(21): 7123-7133, 2004.

MIROSLAVOVA, N. S. & BUSBY, S. J. Investigations of the modular structure of bacterial promoters. Biochem. Soc. Symp., (73): 1-10, 2006.

MITCHELL, J. E.; ZHENG, D.; BUSBY, S. J. & MINCHIN, S. D. Identification and analysis of 'extended -10' promoters in *Escherichia coli*. Nucleic Acids Res., 31(16): 4689-4695, 2003.

MITTENHUBER, G. An inventory of genes encoding RNA polymerase sigma factors in 31 completely sequenced eubacterial genomes. J. Mol. Microbiol. Biotechnol., 4(1): 77-91, 2002.

MOITINHO-SILVA, L.; HEINECK, B. L.; REOLON, L. A.; PAES, J. A.; KLEIN, C. S.; REBELATTO, R.; SCHRANK, I. S.; ZAHA, A. & FERREIRA, H. B. *Mycoplasma hyopneumoniae* type I signal peptidase: expression and evaluation of its diagnostic potential. Vet. Microbiol, 154(3-4): 282-291, 2012.

MOLL, I.; GRILL, S.; GUALERZI, C. O. & BLASI, U. Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control. Mol. Microbiol, 43(1): 239-246, 2002.

MOONEY, R. A.; DARST, S. A. & LANDICK, R. Sigma and RNA polymerase: an on-again, off-again relationship? Mol. Cell, 20(3): 335-345, 2005.

MORAN, N. A. & PLAGUE, G. R. Genomic changes following host restriction in bacteria. Curr. Opin. Genet. Dev., 14(6): 627-633, 2004.

MUSATOVOVA, O.; DHANDAYUTHAPANI, S. & BASEMAN, J. B. Transcriptional starts for cytadherence-related operons of *Mycoplasma genitalium*. FEMS Microbiol. Lett., 229(1): 73-81, 2003.

MUTO, A. & OSAWA, S. The guanine and cytosine content of genomic DNA and bacterial evolution. Proc. Natl. Acad. Sci. U.S.A, 84(1): 166-169, 1987.

MÜNCH, R., KLEIN, J. & JAHN, D. Prediction and Analysis of Gene Regulatory Networks in Prokaryotic Genomes In: Yang, N., editors. Systems and Computational Biology - Molecular and Cellular Experimental Systems. Rijeka, Croatia: InTech, 2011. p. 149-162.

National Center of Biotechnology Information. Complete Microbial Genomes. Disponível em: <http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi?view=1>. Acesso em: 19/01/2012

NICKELS, B. E.; MUKHOPADHYAY, J.; GARRITY, S. J.; EBRIGHT, R. H. & HOCHSCHILD, A. The sigma 70 subunit of RNA polymerase mediates a promoter-proximal pause at the lac promoter. Nat. Struct. Mol. Biol., 11(6): 544-550, 2004.

NISHIDA, K.; FRITH, M. C. & NAKAI, K. Pseudocounts for transcription factor binding sites. Nucleic Acids Res., 37(3): 939-944, 2009.

NOORMOHAMMADI, A. H.; MARKHAM, P. F.; KANCI, A.; WHITHEAR, K. G. & BROWNING, G. F. A novel mechanism for control of antigenic variation in the haemagglutinin gene family of *Mycoplasma synoviae*. Mol. Microbiol., 35(4): 911-923, 2000.

ONEAL, M. J.; SCHAFER, E. R.; MADSEN, M. L. & MINION, F. C. Global transcriptional analysis of *Mycoplasma hyopneumoniae* following exposure to norepinephrine. Microbiology, 154(Pt 9): 2581-2588, 2008.

ORSINI, G.; IGONET, S.; PENE, C.; SCLAVI, B.; BUCKLE, M.; UZAN, M. & KOLB, A. Phage T4 early promoters are resistant to inhibition by the anti-sigma factor AsiA. Mol. Microbiol, 52(4): 1013-1028, 2004.

OSAWA, S.; JUKES, T. H.; WATANABE, K. & MUTO, A. Recent evidence for evolution of the genetic code. Microbiol. Rev., 56(1): 229-264, 1992.

ÖSTERBERG, S.; DEL PESO-SANTOS, T. & SHINGLER, V. Regulation of alternative sigma factor use. Annu. Rev. Microbiol., 6537-55, 2011.

PETERSEN, L.; LARSEN, T. S.; USSERY, D. W.; ON, S. L. & KROGH, A. RpoD promoters in *Campylobacter jejuni* exhibit a strong periodic signal instead of a -35 box. J. Mol. Biol., 326(5): 1361-1372, 2003.

PINTO, P. M.; CHEMALE, G.; DE CASTRO, L. A.; COSTA, A. P.; KICH, J. D.; VAINSTEIN, M. H.; ZAHA, A. & FERREIRA, H. B. Proteomic survey of the pathogenic *Mycoplasma hyopneumoniae* strain 7448 and identification of novel post-translationally modified and antigenic proteins. Vet. Microbiol, 121(1-2): 83-93, 2007.

PITCHER, D. G. & NICHOLAS, R. A. Mycoplasma host specificity: fact or fiction? Vet.J., 170(3): 300-306, 2005.

POLLACK, J. D. Carbohydrate metabolism and energy conservation In: J. Maniloff; R. N. McElhaney; L. R. Finch & J. B. Baseman, editors. Mycoplasmas: Molecular Biology and Pathogenesis. Washington, DC: Am. Soc. Microbiol., 1992. p. 181-200.

QIU, P. Recent advances in computational promoter analysis in understanding the transcriptional regulatory network. Biochem. Biophys. Res. Commun., 309(3): 495-501, 2003.

RANGANNAN, V. & BANSAL, M. Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition. Mol. Biosyst., 5(12): 1758-1769, 2009.

RAZIN, S.; YOGEV, D. & NAOT, Y. Molecular biology and pathogenicity of mycoplasmas. Microbiol. Mol. Biol. Rev., 62(4): 1094-1156, 1998.

RAZIN, S. The Genus *Mycoplasma* and Related Genera (Class Mollicutes) In: Dworkin, M.; Falkow, S.; Rosenberg, E.; Schleifer, K. & Stackbrandt, E., editors. The Prokaryotes. New York: Springer, 2006. p. 836-904.

RENAUDIN, J.; MARAIS, A.; VERDIN, E.; DURET, S.; FOISSAC, X.; LAIGRET, F. & BOVE, J. M. Integrative and free *Spiroplasma citri* oriC plasmids: expression of the *Spiroplasma phoeniceum* spiralin in *Spiroplasma citri*. J. Bacteriol., 177(10): 2870-2877, 1995.

REOLON, L. A. Caracterização funcional de promotores de *Mycoplasma hyopneumoniae*. Monografia (Título de Biomédico) - Faculdade de Biomedicina, UFRGS, Porto Alegre, 2007.

REZNIKOFF, W., BERTRAND, K., DONNELLY, C., KREBS, M., MAQUAT, L., PETERSON, M., WRAY, L., YIN, J. & YU, X. Complex promoters. In: Reznikoff, W.S.; Burgess, R.R.; Dahlberg, J.E.; Gross, C.A.; Record, M.T. & Wickens, M.P., editors. RNA polymerase and the regulation of transcription. New York: Elsevier, 1987. p. 105-113.

ROBISON, K.; MCGUIRE, A. M. & CHURCH, G. M. A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K-12 genome. J. Mol. Biol., 284(2): 241-254, 1998.

ROSS, R. F. Mycoplasmal diseases In: Straw, B. E., editors. Diseases of swine. Iowa: Iowa State University, 1999. p. 495-509.

SAMBROOK, J. & RUSSELL, D. W. Molecular Cloning A Laboratory Manual. New York: Cold Spring Harbor Laboratory Press, 2001.

SASAKI, Y.; ISHIKAWA, J.; YAMASHITA, A.; OSHIMA, K.; KENRI, T.; FURUYA, K.; YOSHINO, C.; HORINO, A.; SHIBA, T.; SASAKI, T. & HATTORI, M. The complete genomic sequence of *Mycoplasma penetrans*, an intracellular bacterial pathogen in humans. Nucleic Acids Res., 30(23): 5293-5300, 2002.

SCHAFER, E. R.; ONEAL, M. J.; MADSEN, M. L. & MINION, F. C. Global transcriptional analysis of *Mycoplasma hyopneumoniae* following exposure to hydrogen peroxide. Microbiology, 153(Pt 11): 3785-3790, 2007.

SCHNEIDER, T. D. & STEPHENS, R. M. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res., 18(20): 6097-6100, 1990.

SHULTZABERGER, R. K.; CHEN, Z.; LEWIS, K. A. & SCHNEIDER, T. D. Anatomy of *Escherichia coli* sigma70 promoters. Nucleic Acids Res., 35(3): 771-788, 2007.

SINOQUET, C.; DEMEY, S. & BRAUN, F. Large-scale computational and statistical analyses of high transcription potentialities in 32 prokaryotic genomes. Nucleic Acids Res., 36(10): 3332-3340, 2008.

SIQUEIRA, F. M.; SCHRANK, A. & SCHRANK, I. S. *Mycoplasma hyopneumoniae* transcription unit organization: genome survey and prediction. DNA Res., 18(6): 413-422, 2011.

SOBESTIANSKY, J.; BARCELLOS, D. E. S. N.; MORES, N.; CARVALHO, L. F. & OLIVEIRA, S. J. Clínica e Patologia Suína. Goiânia: Art 3 impressos especiais, 1999.

STADEN, R.; BEAL, K. F. & BONFIELD, J. K. The Staden package, 1998. Methods Mol. Biol., 132115-130, 2000.

TAMURA, K.; PETERSON, D.; PETERSON, N.; STECHER, G.; NEI, M. & KUMAR, S. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol. Biol. Evol., 28(10): 2731-2739, 2011.

THACKER, E. L. Mycoplasmal disease In: Straw, B. E.; Zimmermann, J. J.; D'Allaire, S. & Taylor, D. J, editors. Diseases of Swine. Ames: Iowa State University Press, 2006. p. 701-717.

THIJS, G.; LESCOT, M.; MARCHAL, K.; ROMBAUTS, S.; DE, M. B.; ROUZE, P. & MOREAU, Y. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. Bioinformatics, 17(12): 1113-1122, 2001.

THOMAS-CHOLLIER, M.; SAND, O.; TURATSINZE, J. V.; JANKY, R.; DEFRANCE, M.; VERVISCH, E.; BROHEE, S. & VAN, H. J. RSAT: regulatory sequence analysis tools. Nucleic Acids Res., 36 (Web Server issue): W119-W127, 2008.

TRUNK, K.; BENKERT, B.; QUACK, N.; MUNCH, R.; SCHEER, M.; GARBE, J.; JANSCH, L.; TROST, M.; WEHLAND, J.; BUER, J.; JAHN, M.; SCHOBERT, M. & JAHN, D. Anaerobic adaptation in *Pseudomonas aeruginosa*: definition of the Anr and Dnr regulons. Environ. Microbiol., 12(6): 1719-1733, 2010.

TURATSINZE, J. V.; THOMAS-CHOLLIER, M.; DEFRANCE, M. & VAN, H. J. Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat. Protoc., 3(10): 1578-1588, 2008.

TURNBOUGH, C. L., JR. Regulation of gene expression by reiterative transcription. Curr. Opin. Microbiol., 14(2): 142-147, 2011.

VALIATI, J. Redes neurais aplicadas ao reconhecimento de regiões promotoras na família Mycoplasmataceae. Tese (Doutorado em Ciência da Computação) - Programa de Pós-Graduação em Computação, UFRGS, Porto Alegre, 2006.

VAN HELDEN, J. Regulatory sequence analysis tools. Nucleic Acids Res., 31(13): 3593-3596, 2003.

VAN HIJUM, S. A.; MEDEMA, M. H. & KUIPERS, O. P. Mechanisms and evolution of control logic in prokaryotic transcriptional regulation. Microbiol. Mol. Biol. Rev., 73(3): 481-509, 2009.

VASCONCELOS, A. T.; FERREIRA, H. B.; BIZARRO, C. V.; BONATTO, S. L.; CARVALHO, M. O.; PINTO, P. M.; ALMEIDA, D. F.; ALMEIDA, L. G.; ALMEIDA, R.; VES-FILHO, L.; ASSUNCAO, E. N.; AZEVEDO, V. A.; BOGO, M. R.; BRIGIDO, M. M.; BROCCHI, M.; BURITY, H. A.; CAMARGO, A. A.; CAMARGO, S. S.; CAREPO, M. S.; CARRARO, D. M.; DE MATTOS CASCARDO, J. C.; CASTRO, L. A.; CAVALCANTI, G.; CHEMALE, G.; COLLEVATTI, R. G.; CUNHA, C. W.; DALLAGIOVANNA, B.; DAMBROS, B. P.; DELLAGOSTIN, O. A.; FALCAO, C.; FANTINATTI-GARBOGGINI, F.; FELIPE, M. S.; FIORENTIN, L.; FRANCO, G. R.; FREITAS, N. S.; FRIAS, D.; GRANGEIRO, T. B.; GRISARD, E. C.; GUIMARAES, C. T.; HUNGRIA, M.; JARDIM, S. N.; KRIEGER, M. A.; LAURINO, J. P.; LIMA, L. F.; LOPES, M. I.; LORETO, E. L.; MADEIRA, H. M.; MANFIO, G. P.; MARANHAO, A. Q.; MARTINKOVICS, C. T.; MEDEIROS, S. R.; MOREIRA, M. A.; NEIVA, M.; RAMALHO-NETO, C. E.; NICOLAS, M. F.; OLIVEIRA, S. C.; PAIXAO, R. F.; PEDROSA, F. O.; PENA, S. D.; PEREIRA, M.; PEREIRA-FERRARI, L.; PIFFER, I.; PINTO, L. S.; POTRICH, D. P.; SALIM, A. C.; SANTOS, F. R.; SCHMITT, R.; SCHNEIDER, M. P.; SCHRANK, A.; SCHRANK, I. S.; SCHUCK, A. F.; SEUANEZ, H. N.; SILVA, D. W.; SILVA, R.; SILVA, S. C.; SOARES, C. M.; SOUZA, K. R.; SOUZA, R. C.; STAATS, C. C.; STEFFENS, M. B.; TEIXEIRA, S. M.; URMENYI, T. P.; VAINSTEIN, M. H.; ZUCCHERATO, L. W.; SIMPSON, A. J. & ZAHA, A. Swine and poultry pathogens: the complete genome sequences of two strains of *Mycoplasma hyopneumoniae* and a strain of *Mycoplasma synoviae*. J. Bacteriol., 187(16): 5568-5577, 2005.

VOGEL, J.; AXMANN, I. M.; HERZEL, H. & HESS, W. R. Experimental and computational analysis of transcriptional start sites in the cyanobacterium *Prochlorococcus* MED4. Nucleic Acids Res., 31(11): 2890-2899, 2003.

VON HIPPEL, P. H. An integrated model of the transcription complex in elongation, termination, and editing. Science, 281(5377): 660-665, 1998.

VOSKUIL, M. I.; VOEPEL, K. & CHAMBLISS, G. H. The -16 region, a vital sequence for the utilization of a promoter in *Bacillus subtilis* and *Escherichia coli*. Mol. Microbiol., 17(2): 271-279, 1995.

VOSKUIL, M. I. & CHAMBLISS, G. H. The -16 region of *Bacillus subtilis* and other gram-positive bacterial promoters. Nucleic Acids Res., 26(15): 3584-3590, 1998.

VOSKUIL, M. I. & CHAMBLISS, G. H. The TRTGn motif stabilizes the transcription initiation open complex. J. Mol. Biol., 322(3): 521-532, 2002.

WALDO, R. H.; POPHAM, P. L.; ROMERO-ARROYO, C. E.; MOTHERSHED, E. A.; LEE, K. K. & KRAUSE, D. C. Transcriptional analysis of the hmw gene cluster of *Mycoplasma pneumoniae*. J. Bacteriol., 181(16): 4978-4985, 1999.

WALKER, K. A. & OSUNA, R. Factors affecting start site selection at the *Escherichia coli* fis promoter. J. Bacteriol., 184(17): 4783-4791, 2002.

WANG, H. & BENHAM, C. J. Promoter prediction and annotation of microbial genomes based on DNA sequence and structural responses to superhelical stress. BMC Bioinformatics, 7248-2006.

WANG, L.; WANG, F. F. & QIAN, W. Evolutionary rewiring and reprogramming of bacterial transcription regulation. J. Genet. Genomics, 38(7): 279-288, 2011.

WASHIO, T.; SASAYAMA, J. & TOMITA, M. Analysis of complete genomes suggests that many prokaryotes do not rely on hairpin formation in transcription termination. Nucleic Acids Res., 26(23): 5456-5463, 1998.

WASSERMAN, W. W. & SANDELIN, A. Applied bioinformatics for the identification of regulatory elements. Nat. Rev. Genet., 5(4): 276-287, 2004.

WEBER, S. S. Fator σ de *Mycoplasma hyopneumoniae*: Mutagênese, Clonagem e Expressão. Dissertação (Mestre em Biologia Celular e Molecular) - Programa de Pós-graduação em Biologia Celular e Molecular, UFRGS, Porto Alegre, 2007.

WEINER, J.; HERRMANN, R. & BROWNING, G. F. Transcription in *Mycoplasma pneumoniae*. Nucleic Acids Res., 28(22): 4488-4496, 2000.

WEINER, J.; ZIMMERMAN, C. U.; GOHLMANN, H. W. & HERRMANN, R. Transcription profiles of the bacterium *Mycoplasma pneumoniae* grown at different temperatures. Nucleic Acids Res., 31(21): 6306-6320, 2003.

WOESE, C. R. Bacterial evolution. Microbiol.Rev., 51(2): 221-271, 1987.

WOLF, M.; MULLER, T.; DANDEKAR, T. & POLLACK, J. D. Phylogeny of Firmicutes with special reference to Mycoplasma (Mollicutes) as inferred from phosphoglycerate kinase amino acid sequence data. Int. J. Syst. Evol. Microbiol., 54(Pt 3): 871-875, 2004.

WOSTEN, M. M. Eubacterial sigma-factors. FEMS Microbiol. Rev., 22(3): 127-150, 1998a.

WOSTEN, M. M.; BOEVE, M.; KOOT, M. G.; VAN NUENEN, A. C. & VAN DER ZEIJST, B. A. Identification of *Campylobacter jejuni* promoter sequences. J. Bacteriol., 180(3): 594-599, 1998b.

YAMAO, F.; MUTO, A.; KAWAUCHI, Y.; IWAMI, M.; IWAGAMI, S.; AZUMI, Y. & OSAWA, S. UGA is read as tryptophan in *Mycoplasma capricolum*. Proc. Natl. Acad. Sci. U.S.A, 82(8): 2306-2309, 1985.

YOGEV, D.; ROSENGARTEN, R.; WATSON-MCKOWN, R. & WISE, K. S. Molecular basis of Mycoplasma surface antigenic variation: a novel set of divergent genes undergo spontaneous mutation of periodic coding regions and 5' regulatory sequences. EMBO J., 10(13): 4069-4079, 1991.

YUS, E.; MAIER, T.; MICHALODIMITRAKIS, K.; VAN, N., V; YAMADA, T.; CHEN, W. H.; WODKE, J. A.; GÜELL, M.; MARTINEZ, S.; BOURGEOIS, R.; KUHNER, S.; RAINERI, E.; LETUNIC, I.; KALININA, O. V.; RODE, M.; HERRMANN, R.; GUTIERREZ-GALLEGO, R.; RUSSELL, R. B.; GAVIN, A. C.; BORK, P. & SERRANO, L. Impact of genome reduction on bacterial metabolism and its regulation. Science, 326(5957): 1263-1268, 2009.

ZHANG, W. & BASEMAN, J. B. Transcriptional regulation of MG_149, an osmoinducible lipoprotein gene from *Mycoplasma genitalium*. Mol. Microbiol., 81(2): 327-339, 2011.

ZIELINSKI, G. C.; YOUNG, T.; ROSS, R. F. & ROSENBUSCH, R. F. Adherence of *Mycoplasma hyopneumoniae* to cell monolayers. Am. J. Vet. Res., 51(3): 339-343, 1990.

8. ANEXOS

8.1 *Primers* gene-específicos utilizados no 5' RLM-RACE

(continua)

Locus	Gene	Primer	Seqüência 5' – 3'	Posição Genômica
MHP7448_0026	*sipS*	sipS_outer	CCAAAAACCTAAATCAGTAAAATTAGCA	33175 – 33202
		sipS_inner	TCGCTGAGGTTTTTTGACATTGTTA	33034 – 33058
MHP7448_0039	*recA*	recA_outer	ATTCCTCCCTGTTTTTGGACTTCA	53094 – 53117
		recA_inner	AATGGCATGGAGACTAATTGTGGTT	53120 – 53144
MHP7448_0040	*licA*	licA_outer	GGAGGAAAATCAAGGTTGGAATTATG	53842 – 53867
		licA_inner	CCACTCCCAAATAATTTTTTCATTAGA	53746 – 53772
MHP7448_0066	*uvrC*	uvrC_outer	AAGATATTTTTTGACTTTTTGCTGCGT	86837 – 86863
		uvrC_inner	GGTTTCGTTTTGCAGAATATCATATTT	87004 – 87030
MHP7448_0067	*dnak*	dnaK_outer	ACGTTGGGCATTGTCAAAATAAGCA	88219 – 88195
		dnaK_inner	AGCGATTGCTTCTGGGTTAGTTTCA	88015 – 88039
MHP7448_0101	*clpB*	clpB_outer	GAAGCACCTGCAAGAATTAAGGTTAGA	136525 – 136551
		clpB_inner	GTTCCCCGACAAGGACAGGATT	136652 – 136673
MHP7448_0161	*deoB*	deoB_outer	TGGGCTAGGCTTAACAGTTTCGA	199699 – 199721
		deoB_inner	TTTCAATTTTGGCGATATTTCCTATC	199571 – 199596
MHP7448_0195	*rpsJ*	rpsJ_outer	GCTTACCTCGATTTCTATTCAAACC	220232 – 220256
		rpsJ_inner	AATATAGTTGAGTTGCTCC	220553 – 220571
MHP7448_0198	0198	P97c_outer	GCCTTGGTATATTCAGGATCTTGAAA	222281 – 222306
		P97c_inner	CAATCTTTCGTGGACTTTCTGATCTG	222070 – 222095
MHP7448_0224	*glyA*	glyA_outer	GCGTTTGCACTTGATCCAGAATAAG	268143 – 268167
		glyA_inner	GCCTTCACCATATTTGTTACTAAGACTTG	268277 – 268305
MHP7448_0225	0225	0225_outer	ACTTTTTCAACCTCTGCAACTGCTTCT	269238 – 269264
		0225_inner	GTTTTTGCCTGAGCTACAATTTTGTCT	269079 – 269105
MHP7448_0241	*secD*	secD_outer	TTCTAATCCGACCTTCGCCTTCA	288011 – 288033
		secD_inner	TGCTGCATCAGCTTCTTTTACAACTT	287932 – 287957
MHP7448_0272	0272	P97_outer	GGCTGAAAAAATCCGAATATGCA	331265 – 331287
		P97_inner	TGAATCTTTTGAGAATTTAACCAAATCA	331395 – 331422
MHP7448_0279	0279	0279_outer	TCTCGATACTGTGTACATAAAGCAGGTTA	338970 – 338998
		0279_inner	GGTTGCTGCAAGTTCATTGATAGTTT	339145 – 339170
MHP7448_0359	*glpK*	glpK_outer	TGATTTGTTAGTCCAAGGGCAACA	450278 – 450301
		glpK_inner	CCTTTGTTTTTGCTGCCTGCATT	450324 – 450346
MHP7448_0360	0360	P37_outer	ATCATCAATGCCGACAAAATCGAA	451191 – 451214
		P37_inner	CAAATCAACTTTGACCAAGACCAAGA	451062 – 451077
MHP7448_0427	*efp*	efp_outer	CTCGATTGCTACTTGTTCATAAGTGGA	541040 – 541066
		efp_inner	CAATATGAGCTTTTTCGACCCTTTCT	550955 – 550980
MHP7448_0454	*acpD*	acpD_outer	AAGTACGATCCGCAAGACAGATTG	599346 – 599369
		acpD_inner	GCGATGAGATGATTACTTTTTGAACAT	599271 – 599297
MHP7448_0490	*pgk*	pgk_outer	TAGGACCACGGTTTTCAGGGATAA	640551 – 640574
		pgk_inner	GCCAAGATGGGATAAAATCACTAATTT	640436 – 640462
MHP7448_0505	0505	lip_outer	TGTGTCCCATCATTTTGAGTAACCA	670037 – 670061
		lip_inner	TTGTGAGCTAGTCAATGGCGATGT	670231 – 670254

(conclusão)

Locus	Gene	Primer	Seqüência 5' – 3'	Posição Genômica
MHP7448_0513	0513	46K_outer	ATCCTGGGACATAAACAGCATT	681010 – 681031
		46K_inner	GTAATTGTTGAAGTTGCTGCCTCTGT	680521 – 680546
MHP7448_0521	*pepF*	pepF_outer	GCCTCAAATTCAGACATCAAGAAACTA	691099 – 691125
		pepF_inner	TTATCCCAAAATTAGCTTCTAGTTCCTT	691211 – 691238
MHP7448_0527	*deoC*	deoC_outer	GTGCCGATTTCCTTCTTAATTTGATT	691211 – 691238
		deoC_inner	TGCACTAATTTGACTTCCTAAAGGGA	703808 – 703833
MHP7448_0528	*gyrA*	gyrA_outer	AACTTGTTCCCGAAGTAATTCCGAGTTCA	704733 – 704761
		gyrA_inner	GCGAAACAATAACAGACATTGAATAATCA	704640 – 704668
MHP7448_0535	*pyrH*	pyrH_outer	TTTCCCTGACAAAATACTCACAAATTC	715824 – 715850
		pyrH_inner	GTTCCAAGCATCCCGATATAATCAG	715937 – 715961
MHP7448_0545	*ktrA*	ktrA_outer	TCTGAAGTGGCTACTATAATCGTGTCAA	727051 – 727078
		ktrA_inner	GCGTTAATATCACTTGCATCCATAATAA	727108 – 727135
MHP7448_0546	*ktrB*	ktrB_outer	CAATTCCGCCAAGGGAGATCAGAA	727846 – 727869
		ktrB_inner	GGTATCGCTAAAGGCAGAAACTGAAG	727759 – 727784
MHP7448_0586	*nusA*	nusA_outer	GGATCAGTTTCTTTTGCCTTTGATAGT	777012 – 777038
		nusA_inner	CGCATCTGGGTCAATTTTCTTTGT	777173 – 777196
MHP7448_0619	*rplJ*	rplJ_outer	GACTTTGAAGACCAGAAGCTAAC	825975 – 825997
		rplJ_inner	TGCATCTGTTGTTCCAAAAGCGAAT	826144 – 826168
MHP7448_0622	*dam*	dam_outer	GCTATTTTCAGGAATTTCAGACCTTAATT	833199 – 833227
		dam_inner	TCATTAACAATAGCATTTTCAGGTTGTAA	833338 – 833366
MHP7448_0647	*leuS*	leuS_outer	TCAGAGGTGTTAATTTCCTTATCCCAA	868099 – 868125
		leuS_inner	GGATGAAGCACGTCAAAACCATT	868258 – 868280
MHP7448_0648	*uvrB*	uvrB_outer	CATCAAGGCGTTTAGGGCTGACAT	869061 – 869084
		uvrB_inner	TTTCCAGATCCTGTTACTCCAAGCA	868801 – 868825
MHP7448_0654	*prsA*	prsA_outer	GGCTGTCTTCCTGATGATTTTCGAT	875123 – 875147
		prsA_inner	GGCATAACCTCTTTTTAACGAGTCAAT	875050 – 875076
MHP7448_0663	0663	p146_outer	TTTGGGGCAAAGGTGACCTCAA	894044 – 894065
		p146_inner	AAGGCAACATTTTGAACTTTTGAAACT	894209 – 894235

8.2 Seqüências promotoras alinhadas usadas na criação dos *logos* de seqüência

(continua)

Bactéria / seqüências alinhadas	Nº de seqüências	Referência
Sinohrizobium meliloti	25 sítios	MacLellan *et al*. (2006)

```
>incA1                  AAAAATGCTTGACACTGATT-CGCGGAAGTGGGATTCTCTAGTTGCTACA
>incA2                  AAGGACTCTTGACTGTGATT-CGTGGAAATGCGATTCTCGATCTTGCTAC
>pckA                   TGGCTATCTTGTCTTGGGTC-AGCCTTGCCGGTATGTTCCGACGAAATTC
>ntrA                   GTCCACGCTTGACCAAATTC-CAGTAATAAGCAATTTTTGGGCCAACTAA
>hemA                   GCAATTGCTTGACTTCGATC-GATGTTCGGGAGAATGAAGTTTTGCCAGC
>hemA                   GTCCGGGGTTGACCACTGAT-CGCTTTGAAGGAAGAAAGGCGACAGGGCA
>nodD1                  GCCGCACCTTGATTCCATTAACTTCAGGGTTCTCTAATAGGACTCTGCAA
>nodD2                  GCTGAAGCTTGATTTCGTTAACTTCAGAGTTCTCTAATGGGAGTGTGAAA
>trpE                   TTTCCCGCTTGCGCGCCATC-GCAAGCCGCGCTAACACTTCCGCCCATGG
>rRNA-16S               AATCGTTGTTGACGTGTTGGAGGGCTGGGGTCTATAAGCCCGATCACTGA
>SMc00029               ATAGGCGCTTGTGAAATCAA-AAATGCTGTTCTATAGAGGCGCCGCTGGT
>repA2                  AGCGCGCATTGATGGCCATAGCCGAATCGGAGTACGTCTTAGTTGCGGAA
>SM_b20587              GAAGTAAATTGACGCTTGCG-AAGAACAGGCGCTTATTCGGCTCGAACAT
>rpmJ                   GCACGCGGTTGACACGATTG-GGTCGACACGGTATGTGCCTGCCTACTTG
>topA                   CAAACCCCTTGACCGCGCCC-GAATCCCTGTCCATGTCGTGGCGTCGGAT
>sigA                   GGAAACGCTTGACGGGATGA-AAAATTCTGGGAATCACCATTTCAAGCAG
>metC                   ATTTTTCCTTGACCCGGCCG-GTGGCCCGTTCGATGAGTCGGCAGCGAAA
>ropB1                  TTTGGCTCTTGAGATTCCTC-ATTTCCTGATCAATTTCGGGTCAACCGGG
>trpS                   CATGCCTCTTGATTGCGCCC-GGCCCGCCGTGCATAAGCGCGGCGGATAT
>rpsT                   ATTGCGCGTTGACGTGCGGG-CGGTTTCCCTTTATACGCCGCCCTCAGTT
>tRNA-MET_CAT           GACATTGCTTGACACTAGCCGGAACAGCGCCTTATAAGCCGCCCGACCAG
>tRNA-TYR_GTA           TCGCCGTGTTGACAGGACGA-ATGCAGGCGGTCATATACCCGGCGCAGAT
>tRNA-Gly (SMo02206)    CATTGTGCTTGACAGATCACGGCTGTTGAGGAAATTTCCGGTTCTGTCGC
>secE                   ATTGGCTCTTGTGGTTTTCG-CAATCGGTCTTTATGTAGGGTCCAAACAG
>trkH                   TTCACCCGTTGAGTTTGCCG-ATGGCAGTCGCTATTGCACATTGCCGCGC
```

Escherichia coli	59 sítios	Hawley and McClure (1983)

```
>araBAD                 CTACCTGACGCTTTT---TATCGCAACTCTCTACTGTTTCTCCATACCCG
>araC                   AATGTGGACTTTTCT----GCCGTGATTATAGACACTTTTGTTACGCGTT
>galP1                  CCATGTCACACTTTTCGCATCTTTGTTATGCTATGGTTATTTCATACCAT
>galP2                  TTCCATGTCACACTT---TTCGCATCTTTGTTATGCTATGGTTATTTCAT
>lacP1                  AGGCTTTACACTTTA---TGCTTCCGGCTCGTATGTTGTGTGGAATTGTG
>lacP2                  TTAGCTCACTCATTA--GGCACCCCAGGCTTTACACTTTATGCTTCCGGC
>lacI                   AATGGCGCAAAACCT----TTCGCGGTATGGCATGATAGCGCCCGGAAGA
>malEFG                 AGGATGGAAAGAGGT-----TGCCGTATAAAGAAACTAGAGTCCGTTTAG
>malK                   AGGATTTAAGCCATC-----TCCTGATGACGCATAGTCAGCCCATCATGA
>malt                   TCGCTTGCATTAGAA----AGGTTTCTGGCCGACCTTATAACCATTAATT
>tnaA                   AGAATAGACAAAAAC-----TCTGAGTGTAATAATGTAGCCTCGTGTCTT
>deoP1                  TTATTCGAACATCGA---TCTCGTCTTGTGTTAGAATTCTAACATACGGT
>deoP2                  TGTATCGAAGTGTGT--TGCGGAGTAGATGTTAGAATACTAACAAACTCG
>trp                    GCTGTTGACAATTAA----TCATCGAACTAGTTAACTAGTACGCAAGTTC
>trpR                   GTTACTGATCCGCAC---GTTTATGATATGCTATCGTACTCTTTAGCGAG
>aroH                   ACTAGTGCATTAGCT-----TATTTTTTTGTTATCATGCTAACCACCCGG
>trpP2                  AACCGTGACATTTTA----ACACGTTTGTTACAAGGTAAAGGCGACGCCG
>his                    GTTCTTGCTTTCTAA---CGTGAAAGTGGTTTAGGTTAAAAGACATCAGT
>hisA                   CTAATTAATAAATAG-TTAATTAACGCTCATCATTGTACAATGAACTGTA
>leu                    AGGGTTGACATCCGT----TTTTGTATCCAGTAACTCTAAAAGCATATCG
>ilvGEDA                TATCTTGTACTATTT---ACAAAACCTATGGTAACTCTTTAGGCATTCCT
>argCBH                 ATTGTTGACACACCT------CTGGTCATGATAGTATCAATATTCATGCA
>thr                    TTTATTGACTTAGGT----CACTAAATACTTTAACCAATATAGGCATAGC
>bioA                   AAACGTGTtTTTTGT---TGTTAATTCGGTGTAGACTTGTAAACCTAAAT
>bioB                   CGACTTGTAAACCAA----ATTGAAAAGATTTAGGTTTACAAGTCTACAC
>fol                    CCAGTCGACGACGGT----TTACGCTTTACGTATAGTGGCGACAATTTTT
>uvrB P1                TTTGTTGGCATAATT----AAGTACGACGAGTAAAATTACATACCTGCCC
>uvrB P2                TATGGTGATGAACTG---TTTTTTTATCCAGTATAATTTGTTGGCATAAT
>uvrB P3                ACTATTCCTGTGGAT----AACCATGTGTATTAGAGTTAGAAAACACGAG
>recA                   ACACTTGATACTGTA-----TGAGCATACAGTATAATTGCTTCAACAGAA
>lexA                   ATGGTTCCAAAATCG----CCTTTTGCTGTATATACTCACAGCATAACTG
>ampC                   ACAGTTGTCACGCTG-----ATTGGTGTCGTTACAATCTAACGCATCGCC
>lpp                    AATATTCTCAACATA---AAAAACTTTGTGTAATACTTGTAACGCTACAT
>hisJ                   TGCTTTGCCTTGTCG---GCCTGATTAATGGCACGATAGTCGCATCGGAT
>Pori-r                 TCTGTATACTTATTT---GAGTAAATTAACCCACGATCCCAGCCATTCTT
```

(continuação)

Bactéria / seqüências alinhadas	Nº de seqüências	Referência
Escherichia coli	59 sítios	Hawley and McClure (1983)

```
>Pori-1          GTTTTTGAGTTGTGT----ATAACCCCTCATTCTGATCCCAGCTTATACG
>spot 42 RNA     GTGCTTTCTGAACTG----AACAAAAAAGAGTAAAGTTAGTCGCGTAGGG
>M1 RNA          CGGGGTGACAAGGGC----GCGCAAACCCTCTATACTGCGCGCCGAAGCT
>alaS            GTATTTTACCTTCCC---AGTCAAGAAAACTTATCTTATTCCCACTTTTC
>trpS            GCTATCGATCTCAGC------CAGCCTGATGTAATTTATCAGTCTATAAA
>glnS            ACAGTTGTCAGCCTG----TCCCGCTTATAAGATCATACGCCGTTATACG
>tufB            TTAGTTGCATGAACT----CGCATGTCTCCATAGAATGCGCGCTACTTGA
>tyrT            ACACTTTACAGCGGC-----GCGTCATTTGATATGATGCGCCCCGCTTCC
>leul tRNA       ACTATTGACGAAAAG-----CTGAAAACCACTAGAATGCGCCTCCGTGGT
>supB-E          GAGGTTGACGCTGCA----AGGCTCTATACGCATAATGCGCCCCGCAACG
>rrnAB P1        CCTCTTGTCAGGCCG-----GAATAACTCCCTATAATGCGCCACCACTGA
>rrnG P1         TCGCTTGTCAGGCCG-----GAATAACTCCCTATAATGCGCCACCACTGA
>rrnD P1         ATACTTGTGCAAAAA-----ATTGGGATCCCTATAATGCGCCTCCGTTGA
>rrnE P1         TCTATTGCGGCCTGC-----GGAGAACTCCCTATAATGCGCCTCCATCGA
>rrnX P1         CCGCTTGTCTTCCTG-----AGCCGACTCCCTATAATGCGCCTCCATCGA
>rrnAB P2        ATGCTTGACTCTGTA-----GCGGGAAGGCGTATTATGCACACCCCGCGC
>rrnG P2         ATGCTTGACTCTGTA-----GCGGGAAGGCGTATTATGCACACCGCCGCG
>rrnDEX P2       AGGGTTGACTCTGAA-----AGAGGAAAGCGTAATATACGCCACCTCGCG
>str            TTTCTTGACACCTTT----TCGGCATCGCCCTAAAATTCGGCGTCCTCAT
>spc            TTTCTACCCATATCC----TTGAAGCGGTGTTATAATGCCGCGCCCTCGA
>S10            ACGCTTGCGTTCGGT-----GGTTAAGTATGTATAATGCGCGGGCTTGTC
>rpoA           TTTCTTGCAAAGTTG----GGTTGAGCTGGCTAGATTAGCCAGCCAATCT
>rplJ           TGCCTTTACGTGGGC----GGTGATTTTGTCTACAATCTTACCCCCACGT
>rpoB           TACTGCGACAGGACG-----TCCGTTCTGTGTAAATCGCAATGAAATGGT
```

Bacillus subtilis	142 sítios	Helmann (1995)

```
>abr            AATGATTGACGATTATT---GGAAACCTTGTTATGCTATGAAGgTAAGGA
>acka           TTGACTTGAAAAGCCGA---CATGACAATGTTTAAATGGAAAAgTCAGAT
>acsA           AAAAATTGAGAAGAATA---TGAATATATACTATAATAATTGTgACAACT
>acuA           CCATGTTGAAAACGCTT----TATAATTTGGTATTCTTAAAGAaGGCATG
>ada            CCACCTTATTTTTTTCT---ATTTATGAGGTTATAGTGTAGTTaTCAAGA
>addA           CAGATTGGTCATTTTCG---TCAACATTCGATAAAATATAGaGAGATAAA
>ald            AAACAAAGAATTTTCCA---AAATATCAAGCTACACTAAAAATaTCACAT
>alkA           ACAAAATGAGTAAAGAT----GATTATGTGATAAACTAATTTCaACCAGC
>alsS           CGATATTGGAGGTCAAT--TTCCAAAGAGTGTATAGTGAAACTTaTCACA
>amyLY          CCAAGTTGTTTTGATAG---AGTGATTGTGATAATTTTAAATGTaAGCGT
>ansA           ATGTCTTGCGTATTTTA---TTTTCCGTGTCTATAATTGCATtAAAAGTA
>aprE           ATGGGTCTACTAAAATA-TTATTCCATCTATTACAATAAATTCaCAGAAT
>argC           TACGATTGAATTAATTT----TTATTCATGTTATAATGTTAAATaATTTC
>asd            TTCGGTCATCACCCGCA----TATTCTATGGTAAAATAAAACGtAAAATC
>cdd            TGGGCGTGAAAAAAAGC---GCGCGATTATGTAAAATATAAaGTGATAGC
>citA           AAATATTGATTATTTTT---TAAATATTATATTTACTATAATAaCAGAAA
>citB           TTGATTTTACTATATGT--ATGATTTTGTTTTAATATGAAATTgTGAGAA
>citC           TTGTTCAGACATCAAAA---TTGGGTTACACTTTAAATTGAATgTTAGGA
>citR           CTTTTCTGTTATTATAG---TAAATATAATATTTAAAAAATAaTCAATAT
>citZ           AACATTTAACAAATGTC---TGATTATTGTTTATAATGAGAATaGGCTTA
>comA           ACGACTTGGCACAGGCC----AAGTCTTTTTTATAAAATGGAaaAGAGTG
>comB           ATGTTTAGACAATTTTCGTCAAATTATTTGATATACTTAGGGGTgAAAGC
>comC           CCCAATTGGAGCTCAGG-ATGTCCATTTTGCTACAATCCATACCAgAACT
>comE           TGCTTTTGCAGTACAGA---CGAACGTATGACATACTCGTCTaCACATGA
>comF           ATAGTTGGACAGAAAAT---ATTCATTCAGGCATACTGTTTCgAAAGGAG
>comG           CCTGTTTGATTACCTTT---TCTTCTTTTTCTACAATATGCGTTgAAAGG
>comK           AAAACCTGTCGTTTTAC--AAAAACAGATGATAGATTATTAgTATAAATT
>comQ           ACCTGCTGTCCTTTAAA--TGTCCCATTTAGTAAAATGGAATgGGAGGGG
>cspB           AGCAGTTGTTTTTTCTG---AAGATTACTGGTGGAGTAAAGGTaATTATT
>ctaA           ACATATTCGCTTACACT---TGGAACGTATATAAAATTGCaGCAGTATGT
>dciA           TTATATTGCATTTTTCC---TCTTTTTTTAATATAATTTGTTaGAATATT
>degQ           TTTCGGTGAAAAATGAG---CCGAAAGCAGACACACTATTAgTAACAGAT
>degSU          GTCTTTAGAATTTTTGT---GCGTATTTTGGTATCATAAAGAGTAGaTAG
>divIB          AAGTTCTGACTGAAGCT----GTTCATATGATATACTGTAAGCaAACGAC
>dnaA           AACCATTGCAAGCTCTC---GTTTATTTTGGTATTATATTTGTgTTTTAA
>dnaE P2        AAACAATAGCATCTTTG--TGAAGTTTGTATTATAATAAAAAATTgTGAT
>dnaG           CACAGGCCCTACTATTACTTCTACTATTTTTTATAAATATATaTATTAAT
>dnaK           TATAATTGACATTTTTC---TTGTGGTTTGATACTTTTGTTATaGAATTA
>ffhrpsp        AAAACTTGACAGTGTCA---TTAAAACCGTGTAAACTAAGTTATcGTAAA
>ftsA P1        ATTCATTGTATTGTTGT---TCCGCAAATAATAGAATAGAAATgATCGAA
>ftsA P3        TTAGCTTTTCTGGGAGT---CCATCTTGGTGTAGACTTGTATTTAGCAGG
>gcaD(tms)      TCTCCTTGAAATCAGAA---GATATTTAGGATATATTTTTCTATgGATAA
>glnR           TCGTATTGACACAAAAT---ATAACATCACCTATAATGAAACTAaGTTAA
>glpD           TGGCTTTTAAATAAAGT----AATACTATGGTATAATGGTTACaAGTTAA
```

(continuação)

Bactéria / seqüências alinhadas	Nº de seqüências	Referência
Bacillus subtilis	142 sítios	Helmann (1995)

```
>glpFK      AGTGATTGACACCGCTT---TCATGCACTGATACAATTGCACTaGGTTAA
>glpTQ      ACCCGTTTACATTCGCT---CCGGACTATGATAATTTAAGAAAGCGCTAT
>gltA       ATATATTGTTTATATCG---TTTTGAAAACCTACAATGATTATaGAGTTG
>gltC       GTAGGTTTTCAAAACGA---TATAAACAATATATAATTTAGATCaaAAGA
>gltX       TTGTTTTGATGCCGGGT---CTATTTGGTGGTAGAATAGATTCaTACATT
>gluB       GTTGATAGACAATCATG---AGAAAGATTTTTACAATGAGTTCgaGCTCA
>gntR       AGTGTTTGCATAAAAGA---AATATTCACGTTATCATACTTGTATaCAAG
>groESL     ATCACTTGAAATTGGAA---GGGAGATTCTTTATTATAAGAATtgTGTTA
>gsiA       TTGCGTTCTTCTTTTTC--TTGTAAATATGATAAAATATGACaTaTCTCG
>gsiB       GTTTGTTTAAAAGAATT---GTGAGCGGGAATACAACAACCAaCACCAAT
>guaA       TGGTCTTGACCGCTTAT---CGACGTGTTGTTAGAATTAGTGAaTATTAT
>gutB       GGCTGTTGAACGTGTTA-----AAAAGCAGATAAAATGGGGGCaGTAACA
>gyrB       TTTAATGTGTATAATCA--TAAGTTTATTGATATAATGGAGAatagTGAA
>hbs        AATGCTTGATATGGCTT---TTTATATGTGTTACTCTACATACaGAAATT
>hemA       ATCTGTTCACATTTTGTGAAAGAAACTATGTTATAATTATTATAAATAAT
>hsmB1      GTTTATTGTGTTAATCA-TTACAAAAAATGATAAAATAAAAAAGcAAGAC
>hut        ACCTTTTGACTTCTGCT---GCTGAACCAATTAATATAATACTCaGTTAA
>ilvB       TGCTGTTGACACTGCGT---CCAAAGCGGCGTAATATGAGTTCaaCAAAA
>kinC       TATTTTACTCAATTATT-TGTCGAAGAATGGTACAATAAGTAGaGAAACA
>lon S1     TTCCTTTTTCATATCAG----AAAGAAAGGGTATACTACGAGGAgACTGT
>lon S2     TTTATTGTACAAGGATG—AAAAAGTTGTATTATAATGGTTCATaCTAAA
>lysC       TAATGTTGTCCTTTTAA---ATAAGATCTGATAAAATGTGAACTaaTTTC
>mecA       AGATGTGGACGGAATGG---GTAAAGTGTAGTAAAGTACAATTAaTCGGG
>men        TTTCTTTTTCATTTCTG----AAATTAGGTTTATAATAGGTAAgGCAGGC
>menB       TCACATAGAGAATAAAA---GGAGGTCATCATATGGCTGAATGGaAAACA
>menE       TTTTCAATACAGACATT----TTACCTCGGAGATGATGACATGCTGACAG
>nadB       AACTCTTGAGTTTATTT---TATCCTTGTGTAAATATAGGTGTCaAGACA
>nasA       TGTCTTTTCCAGAAAAT----AATGGTCCTATATCCTTGATTCaGAAAAT
>nasB       ACGTTTAATGCGTTAACAATGCATTGTGACATAATTTTTAATaGGAGAAA
>nifS       CTGTCTTGACACCTATA--TTTACACAAGGATAAAATAAACTCaAGAGTT
>nrgA       TCTTACATGAAAATGTTTTATCAT---TCTTTTTTCTCTATAATGAAGAA
>nusA       TCCATTTGCAATAAAAA---TATGGTTATGGTATAGTTTTATtgGAAATG
>odhA       CTATTTTGTGAACAATC----AAGGTAGAATCAAATTGCAAACAGtGGTA
>orfS       ACATTTTCTGAGCATTT--TCTCTTTTGTTGTATACTGATATTGTACGTT
>pbpD       TAACCTTCTCGTAATCT---CAAAAGAATGGTACGATATGGCTaGAATTT
>pbpE       AATATTTGAAACGTTAG-TAGGTTAGTAACGTACAGAGATATgGGAGGTG
>pbpF       CCTGCTTGCTAGTATAT---CAAAACAATGGTATAAGTTTCTATTgGCGA
>phoA       TCAACAGCCGCATTTAA---CAAAGTTTCCCTAACATGATAAACgGAATA
>ppiB       ACAAGGAGAGACAGGCT----ATAACCTATACATAATCTCCTTTgAAAAT
>ptsXHI     CGTGCTTGTCAGATGAC---AAGTACGGTTGTATGATATAATaTTGTGAA
>purA       TACAATTGACTTTCTGT---TTCTTCACTGATAAACTTGATTTgTTTGAA
>purF       TATCGTTGACATTATCC---ATGTCCGTTGTTAAGATAAACaTGAAATCA
>pyr        AACGGTTGACAGAGGGT---TTCTTTTCTGAAATAATAAACGAAgCTGAA
>rbs        TTTTATTGCTAACTTCG--GATTGTTCATGATAATCTATCTATgTAAACG
>recA       TTTTCTTGGCAAATCCC---TTCAAACAGGGTATAGTATATgtAGTGGTA
>ribG       TTTCATTGCGTACTTTA---AAAAGGATCGCTATAATAACCAATaAGGAC
>rpmH       TGAAGTTGACAATGAAT--AGGTAACGCAAATATAATAAGTAAGACtgtC
>rpsd       AATGGTTGACTTCAAAA---CAAATAAATTATATAATGACCTTTgTGTGA
>rrnA P1    ATGTATTGACTTAGACA---ACTGAAGGTGTTATTCTAATATCGCTGATG
>rrnA P2    AGTTGTTGACAGTAGCG---GCGGTAAATGTTATGATAATAAAGTCGCTT
>rrnB P1    AACTATTGCAATAAATA---AATACAGGTGTTATATTATTAAACgTCGCT
>rrnB P2    AGTTGTTGACAAAAAAG---AAGCTGAATGTTATATTAGTAaAGCTGCTT
>rrnD P1    AGGTGTTGACTCTGATT---CTTGACCGTGTTATATTATTAAACGTCCTG
>rrnD P2    AACATTTGACAAAAGAA---AGTCAAAATGTTATATTAATAAAGTCGCTT
>rrnG       AGTTATTGACTTTGAAG---AAGTGACATTGTATACTAATAAAGTTGCTT
>rrnJ P1    AACTATTGCACTATTAT---TTACTAGGTGGTATATTATTATTCGTTGCC
>rrnJ P2    AGTTATTGACTTCACTG---AGTCAAGGAGTTATAATAATAAAGTCGCTT
>rrnO P1    AACCCTTTACAGTCATA----AAAATTATGGTATAATCATTTCTgTTGTC
>rrnO P2    AAGTATTGACCTAGTTA---ACTAAAAATGTTACTATTAAGTAgTCGCTT
>rrnW       AGTTGTTGACTTTGAAG---AAGTGACGTTGTATACTAATAAAGTTGCTT
>sacB       ACCAGTTGCAATCCAAA---CGAGAGTCTAATAGAATGAGGTCgAAAAGT
>sacXY      TACTATTGCTATACAGC---CATGAACAGCATAAAATGAACGTTaTTACA
>sdh        TTTTCTTGACGCCCTTT---TGAGGGAGGAGTAAAATGAAATTgTCAATA
>sigB       TTTTTTTGTTTTCAAAA---AACATAAACGATATAATAGTgAaaTAACGA
>spa        GGATCTTGATATTTTTTTGATTTTTAGAATGTATAGTAAAAATAgAGTAT
>spo0A      CCCTCTTCACTTCTCAG---AATACATACGGTAAAATATACAAAAgaAGA
>spo0B      ATTGTTTTTCTAACAAAGCCTTCTCTGTTATAATTCATAATaCACACTTA
>spo0E      ATATGTTTACAAATAAA---GTATAATCTGTAATAATGCACAATAaCCCA
>spo0F      ATGCTCAGAAAATGTCG-TAAAGTAGACTATTATAATTAAAgGAAATAGG
>spo0H      TTAAGTTGACGCTTTTT--TGCCCAATACTGTATAATATTTCTaTCTACG
>spoIIE     TTCTTTTGACAAAATCC-TATCTGTGCTCGCTATAATGACAGGCAaCGAA
```

(continuação)

Bactéria / seqüências alinhadas	Nº de seqüências	Referência
Bacillus subtilis	142 sítios	Helmann (1995)
>spoIIG	TTTCCCACAGAGCTTGCTTTATACTTATGAaGCAAGAAGGGGAACAGCGT	
>srfA	TTTCGGTGATAAAAACA---TTTTTTTCATTTAAACTGAACGGTaGAAAG	
>tet	AAAGTTTAATCCTTAGT----CTATATATAATAAGATCATATCaATCAAA	
>thrS	TCGTGTTGATTTTTTGG---ATTGAACAATTTATAATACATaggAGATTA	
>trnS	ACCTCTTGACACTGCAA---ATCAAGGCTGATATAATAAGTCTTgTCTCA	
>trpE	TATCATTGACAAAAAAT--ACTGAATTGTAATACGATAAGAACagCTTAG	
>tsr	TTCTGTTGCCGATTTGT---CGAAAAGTTGGTATCCTAGTTATgGAGAAA	
>tyrS	CGGCGTTGACACAGGAT---TTTATTTATGTTAAAAATGATATaGCTTCA	
>veg	TTTATTTGACAAAAATG---GGCTCGTGTTGTACAATAAATGTaGTGAGG	
>xylR	AGATGTTGAAAAAGTCG---AAAGGATTTTATAATATTAAGTCAAGTTAG	
>f01epr	CGGTTTTGACAGAGAAG------AAATTTGGTATAACCGTGAGCGCAGTG	
>f105	CGACTTTTACAAAATGT---CGTGAATACCATACAATTTAGACATACCTT	
>fspp1epr	AGTGGTTGCCTTTCTAT---GTTTTCTATGTTTTAATAGAATCATAGAGA	
>f01e22	AGGTATTGACTTTCCCT---ACAGGGTGTGTAATAATTTAATTaCAGGCG	
>f01e3	AGTTGTTGACTTTATCT---ACAAGGTGTGGCATAATAATCTTaACAACA	
>f105epr	TAGTATTGTATTTCCGA---CATTCGGATACTATAATTGTGTCATGCCAC	
>f29 A1	AATGTTTGACAACTATT---ACAGAGTATGCTATAATGGTAGTaTCAATG	
>f29 A3	ACAAATCCTTATGTATCAAGGGTTCACGTGGTATAATTAAGTAgTACTAA	
>f29 B2	TCCGATACACACAAAGC-CGTATAAACCGTGTATAATAGGGgtAACCCGC	
>f29 Ec3	CAACGTTTACAAAGTGA---ACAGGAAGTGTTAAACTATATAGagACACA	
>f29 G2	AAGGGTAGACAAACTAT--CGTTTAACATGTTATACTATAATAgAAGTAA	
>f29 G3a	AAGTCTTGCAAAAAGTT--ATACAGGTGTGGTTAAATAGAGAACgTAGAC	
>f29 G3b	AAGTGTTGAAAATTGTC---GAACAGGGTGATATAATAAAagAGTAGAAG	
>f82-129	AGTTGTTGACATTTCTT---CCCATCCATGCTATAATAAAGTCaTAGAGA	
>f82-156	AGTTGTTGACTTTCTCT---ACGAGGTGTGGCATAATAATCTTaACAACA	
>fsp82	ATTGATTGACTTCTGCC---ACCAAGTGTGCAATAATTATAGTAACATCA	
>fsprmtas	AAATGATCAATTTTATTTAGAGTAAAAATAAAATATaTgGAGGTTGTTTA	
Chlamydia trachomatis D	41 sítios	Grech *et al*. (2007)
>CT017	GTCGCAGTTGACTTTTTCCTTT-AAGTCAATAATAATTCCCTCTCTAGAG	
>CT043	CTAATCATTTACTAACAAACCT-GCTTATGCTAGGTTTAAAAAAAACAGT	
>tyrS	GCTTGCCTTGCTATAAAAAGAACAGGATAGATAAGATGTTGCTAGATAAG	
>CT066	ACACTAATTGATTTTTATTTCG-ACTGAACATTAAATCGAAAAAAACAGA	
>rpsA	GGAAATCTTGCCTTTTTTAAGG-TGAATATTTACACTACTCTTTTTGACT	
>groES	AAACCAGTTGCAAAAAAGCGAG-GACTTTGCTATCGTTCTTCCTCTGAAC	
>tRNAAla_1	TCTTTTCTTGTGCTTAAAATCGCTCTTGGATTAAGATGGCGCTTTGTTAC	
>gcp_1	TAACGCCTTGCTTGATTAACAA-TCTCATGATACGATCCTCTCCTTCCAA	
>accA	ATTAAAATTGTTTGCGTGAAACAAAGGTCATTATAATCAAATAGTTGGTT	
>CT266	AGCAAGCTTGACTCTAAATTTC-CTCAAGATTATTTTTTGCCATTGGACG	
>ihfA	AAAAGTCTTGAATCCAAAGGA--TGAATGCATATTATACGCATATATTGC	
>murE	ACAAAGCTTGACAACGAATAT--GTGTATAGTAAACTATTTGAGAAAGCT	
>CT273	ACCATACTTGACTTTTTCCCT--CCCCCGATTATGATTGAGATTGTGAGC	
>clpC	CGAAAAGTTGCATCATTATCAT-AAATGTCGTATATGCTTGAAAAATATT	
>rnc	AGTACTATAGACTTTAAGATTA-TTTCCGCCTATAAAAACCCGATTGACT	
>infA	AAGTTGTTTGACATTTTCTGT--TTAGTCGATATAATCGCTCTCTCGAGT	
>rpsU	GTATCTCTTGAAGCCTAAATAAAAGTGGTGTTACAATCCCCGGTCTCTTG	
>proS	ACAGAGATTGATCTAGAAACAC-TCCTATGCTAAGATGCTCTTCCACACG	
>CT398	AACGTGCTTTACTTCTTGCAGA-AAAATCGGTAAACTTGCCGTTTCGTCT	
>rpsL	TAACCCCTTGCAAACAAAGATATTCTTATTCTATATTTCCCTGTTTGTAA	
>euo	AAACCGCTTGATTAATAAGTTT-TTTGTTGGGAAAATGTTACCTTCTCTT	
>murA	TTTGTTTTTAAAAACAAACAAATGTCTCTTTGTTAATAAGATGTTTTTTT	
>pheT	CAAAATACTTGCTACTATACACGCCACTTCGTAAAATCTACCAAAAAAAA	
>pgsA_1	TCTCTGCTTGCTTTTGGAGTGT-CTATGTTTCATAATATGTGTCATTGCG	
>rplC	TTTCTTATTGTAAAAATCGTCT-TCCTTTGATAATCTGTCCCTTTAATTT	
>CT546	AATTTATTTGGCATTGCTGTTT-TTATTTATTAAAATAAATAAAAAGGTT	
>CT547	AAGAGATTTGACAAATTCTCTTTTTCTTTTTTATGATGACGCTTTGTTAT	
>yscJ	TTCCCGATTGGCACTAATCTCC-CCATTTGCTATGGTGAGTGAAAAGGTG	
>exbB	AAGGATCTTGGTTCTATACAAG-AAATTTGTTAGGATCGTCTAGGAACTC	
>rpsD	AAATCCGTTGTAGAAAATTGGA-AATAGAACTAGAATGCTCTTTTGTATT	
>CT646	GTTTTTCTTGAAAAAGATGTTT-TTATTTTTTAAAATGAGCGCTCTTCAT	
>ompA	TATCAACTTTACGAGAATAAGAAAATTTTGTTATGGTCTCGAGCATTGAA	
>tRNAGly_2	AAGAAGATTGCATAAAAATCCTTGCTTCCAGTACTATATCGGTCTACTTG	
>clpP	AAAACGCTTGACCCAAGAGACA-CTTAAACATAGAATTCATCATTTTGAT	
>CT708	ATTTTCATTGATTTAGCGGAAG-TAAAAAGGTACAAGTAACAGGTCTGTC	
>efp_2	ATTCTTCTGGACAAAGCTTAGAAGAGAACGATTAACATAGATGGAGAAAAG	
>CT768	AAAACCGTTGACGATAATGCAT-TGCCAGAGCAAACTTTGACTACCAATT	
>ybeB	AAAAAGCTCGACCTTATCTTAGATAATCGGGTATTCTCAGGCCAGTTTCA	
>nrdA	TTTTCAATTGCAGGAAACGTTG-CTAGCTTCTATATATGGTATACAAGAG	
>CT837	AAAATATTTGAAAGCTAATTCATTTATAAAATAAACTAGAAGACAATCTT	
>CT619	ACAAACATAGAAAAAAACTTTT-TTAAATAAGAAAATAAAAACATAAAAG	

Bactéria / seqüências alinhadas	Nº de seqüências	Referência
Mycoplasma pneumoniae M129	35 sítios	Weiner *et al*. (2000)

```
>4.5S RNA          AATTAATTGATTAAAAGCAG---TTAAGTGTTAGAATTTAGAGCCGTCAC
>16S RNA           ATTCTTTAAACATAAATAAAAAGTTTTTCTGTATAATCTTCAGGCTGTTG
>10S RNA           AAGGTTTGGTCAACTTGCGT-TGTTAATGGTGGAGATGACGGGAATCGAA
>MP200 RNA         AGCCATTTGAAATAAGTTTG---GGGTTTGTTAGAATAATTGCACCTAGT
>003, MPN152       CAAGGGTTAAACATCTACAA----ACTTTGTTAATATTTCTCTAAATTTC
>015, MPN140       CTACGACAACAACAGTTGCTGTTTGATTCTTTAAACTTAAACAGAATTAG
>058, MPN097       TTTTGCTGGTATGCTTAGCG---CGTTTTTGTAAGATTTGGCTTTCTTTC
>071, MPN084       CATTAATCGGTTTTTGAGTAAATTTATAAAGTACTATTTTCTTTAACTAG
>072, MPN083       CTCAAGTTGATTAAAACTCT----AAAGAGTTAAAATAGAGTAATGGTTA
>102, MPN052       AATACATTGAAAATCAACGA-TCTAATAAGTTACTATTACGCATATTTTA
>250, MPN592       TAAGGGTTCACCTTTCAAAC--TTTCTTTTTTAAAATTAAAGCACACTAT
>251, MPN591       AACCTAAGTACATTTATGCC----AAACTTATATAATTGGGCACTGTTAA
>268, MPN574       ATTTAACTGCTAATTGTTAA----TAATTGGTAGAATATTGACGGTATGA
>282, MPN560       ACGCAATTAAAGCAATTGTT----AATTATTTAAAATTAATCGAATTTAC
>311, MPN531       GCACTCAAGCCATTCGAGTG--CTAATTTTATATAATTGGCCTATTAACA
>311, MPN531       CGAGTGCTAATTTTATATAA--TTGGCCTATTAACAAAAGAAAGGGGATA
>350, MPN491       CAAAAAGGGAAAAACGCCAT----TTTTTTAGATAATTAAGGCAAATTTT
>381, MPN459       TTAATTTCGCAGATTAGCAA-GCTCCCAAATTAGAATAAAATCACTTTTA
>385, MPN455       CATTTATTAGCAATTAAGCG--CTTACAAATTAAAATTGACCCAGTTGAC
>386, MPN454       TAAGTGGGGAATTAAAGCTT---TGAAAAGTTAAAATTTTCCCAATATTT
>391, MPN449       TAATTAATGCGATTTGGCAG---CTGGATATTAAAATTAAGCGCAATGAT
>394, MPN446       GGGGATTTTAATCTTGATTATTAAGATAATTTAAAATTGTCACCACATGA
>396, MPN444       CATCCCCGGTCAAAGAAAGT-GGTTAAATTTTAGAATTATACCAAATTTT
>438, MPN401       CAAATATTGGGAAAAAAAAGTAGATTTAATTTAACTTTAAACACAATTTC
>438, MPN401       ATTTTTTTGGCAAATATTGG-GAAAAAAAAGTAGATTTAATTTAACTTTA
>443, MPN396       ACTAAAGTGGCCATTTTGCT----AAATTTATATAATTTAAGTCAATTGT
>446, MPN393       TOGTTTTAGTTACTATACCCCTAAAAATAAATAAGATTAAACTAACAAAT
>461, MPN376       GAAGGATTGTCAAAAATATT-TTTAAAGTGCTAGAATAAAGCATGAAACG
>528, MPN309       TATTCATTTGCATTTTTTTA----GATAAATTAAAATTAATGGTATAGTG
>533, MPN304       TTTTTCTTGGGCTTTTGTGT----GGACGTGTAACAATTAAAGCAGTAAG
>548, MPN288       TTAGGAATTAAACTTGTTTA---ATCTTTATTAAGATTTCTCCAAATTTT
>555, MPN281       ATTCTTTTTACTTTCTTAAA-AATATTTTTTTAAGATTTTCCATCTTTTT
>564, MPN271       TTATGAGTTAAAAAGTTCAT---TTTTTAAATAACATTAAACCGGAAATT
>564, MPN271       CATTTTTTAAATAACATTAA--ACCGGAAATTTCAATTAGTTTCCTCTTC
>620, MPN212       AAATGGCTTACAACGAACAA-CACCACAAAACACCAATGACGGTGCAAAA
```

Prochlorococcus marinus subsp. *pastoris* str. CCMP1986	25 sítios	Vogel *et al*. (2003)

```
>pcb               TATTAATTCATCCGTTTGATGAAATATAAAGTATTCTCAAAACGTTTAAG
>psbA              ATTCTTTTTCACATTATTGATGATTTTGGACTATATTTATCAAAGCGAAA
>ftsZ              TTTCGGTTATTCACTCAAATTTTTGACAAGTTAATATTTAAGGACTTTCT
>ftsZ              AAGAAAATCCTACATATCAGCTTAGTGAGTTCATAATGCATCCAATACTA
>psbD              ATGTCTATTCCTACTAATTTAAGAATTAATATAGAGTATTATTACCTAAT
>psbD              GATATTTCCCCACTTTACTCTCAATCTTTGTTACAGTCTCATCGTATCCC
>ntcA              TTTTTTAAAAAATATAAGAATTGAAAAATGTTACTGTTGATACAAGTTAT
>ntcA              TGTTGATACAAGTTATTTCTTGCTGTCTTCTTAAGTTTTTTAATAGTAAA
>petE              AACTTCTCGCGATTTAGTTCCAAAAACCTTGTAATATATATAAATAATAA
>petB              AAGTGTTAACACCAAGTAAGTGAATTTTAAGTAAGCTTTCCATATAAAAG
>rpsL              AAAAATGCTTTAGATATGACTCGTAAAAGGTTATGATGTTTTGATATGTA
>groES             TAAAGAGTGGTTTAATTGCTGCTTAACAGATTATGTTATTTACCATACGG
>atpB              AACCCAGTTAACGAATATACCCTCGGGATGGTAATATTTCGCATTATGAG
>kaiB              CAGTCTTGAATGTTTGTTTATCTTCTTATTTTATAGTGTAATTAGTCAAG
>rplU              ATAAACCTTGACTAATTACACTATAAAATAAGAAGATAAACAAACATTCA
>psbH              AATTTGTAAACAATTTATGTATTAACTCTTATACAATAACTAAATAAATT
>psaF              AACTTTGTGCATATTTGATTCGATACTGTTTTATTTTAATAATAATTTCT
>chlN              ATCAATTAGTGCTTAAATCCAAAAGTTTATGCAAGCTTGAAGAATAAGCT
>PMM0131           ACTATTTAATAAAGTCTATAACATTTATTACTATAGTAATTAATACAAGC
>csoS1             AAAATGCTTGACTTATCAGTACGTTATGGACCATTCTTCGGATTGAACAT
>PMM1101           TGAAATCAGATTTATCTTTACCTCCAGATGTTAGATTAGTGATTCGGATC
>ndhC              CACAACTTCTACATTCGAATCTATTTTGTTCTAAATTGGTAATGAGCTCT
>petH              TAAAACACATACAAATTAATATACAAATTGATAATCTCTTAGTATAAATA
>petH              ACGTAACCATTTTTATAAAACACATACAAATTAATATACAAATTGATAAT
>ureE              AGAAACAATTACCTTCCCAAGAAGATTTAGCCATAATTGTTTTCTTGATT
```

Campylobacter jejuni 81-176	21 sítios	Wosten *et al*. (1998b)

```
>p. ORF1, clone 1b7    ATTTTATGATTTTTACAAT--GAAATTTAGTTATAATTGTAGTTTATAAA
>p. ORF1, clone 1g9    ATGTTAAATTTAATTTATC--TTATTTTTGCTATATTAACGCCATAAAAT
>p. ORF1, clone 2A12   AAATTAAGTGATACAAGA---TTAAATTTAGTAAATTTTAATATAATTAT
>p. ORF1, clone 3D8    ATTTATTGTAAAAATGCA----ATTTTCTGCTAAAATTAAACAAAAATTA
>pspA gene             TTATTTAATTTTTGTAAT---ATTTTTATGCAATAATTATTTATATCTTA
```

(conclusão)

Bactéria / seqüências alinhadas	Nº de seqüências	Referência
Campylobacter jejuni 81-176	21 sítios	Wosten *et al*. (1998b)
>icd gene, partial	TACTCTTGCATTTTTTTT----TATTTGTAATATACTTAAAAATGAAAAT	
>5G10	CGTAAAAGTTTTTTGAAGT--ATTTAAATGATATATTTCAAAATCTAGTT	
>p. ORF1, clone 11B4	TAAACAAATTTGACAAAA----AGTGCATTTTTTTACTAAAATCACATTT	
>p. gene gltX-2	TATTTAAGTATAAATATCGC-ATTATATCAATAAAAATTTAAGAAAAGGC	
>p. ORF1, clone 14B7	GTTTTAAAAATAAGCTTTA--TTTTTCTTTATATATGTTAGAATCATGTG	
>metK gene, partial	TTATCAAGCTTTGATAAG---TTTATTTGGATACAATTGTGGTTCATTTT	
>sodB	ATTTTTACTCTTTTTAAT---ATTATTATGGTAATATTTTCAATATCAAA	
>hup	TTTACTTGATATTGTTTT---TTAAATATGCTAAAATTAGGCGTTTCAAT	
>ileS	AAACAAAGTAGAAATTAAAG-CTAATTTGAGTAGAATTTAAGATTATTAA	
>rpsO	AAATTCAGTTTATTTTTTG--CTTTTTAAGATAAAATTACAGTTCTTTAA	
>tigPI	ATTTTAATAATAATGTAAT--ATTGCAATTTTATCATAAATTACTTTAAA	
>ORF3	AATTTAAGTCTTTTTCTAAG-AATTATTTATTATAATTACAGCTATTTCT	
>proA	TATTTACCTTTCAATTCTA--AATTTTTTTGTATAATTTATCAAGTTTTT	
>orf1	TTTACCAGCTCCACGCTT----GTTATTTTTTATCTTTGTCTTTGTATGC	
>glyA	CAAATAAAAAATCGATTC----GTGATGTGATATTGTTTCCTGCAATGCG	
>lysS	ACAGCAAGCCTTGATTGCAAAAGAACATGGTTTTAAACTTACAGGGCATT	
Mycoplasma hyopneumoniae 7448	23 sítios	Este trabalho
>sipS	AAAATCAAAAATTAAAATTGTTTTTTTATGATAAAATATCAAAGAAATAA	
>recA	TAAATTTTTCCTTTTTTTATTAAAATGTTTAAAAAATATTAAATTAGAAT	
>licA	TAATTTTTATTTTAAAATTTGAAAAATATAATAAAATTTCCAGTATGAAA	
>uvrC	ACTTCAAGATTTAATTATACCAATTTTTTGTTAAAATTATAATAATTATC	
>clpB	ACTCTTTACTTTTAAGTGCCAAAAAATATGTTATAATTTATTTTGTTAAG	
>rpsJ	TAAAAAAATTATTGAATTTTTATTTTTTGTGTATAATTTAATCTTACCGT	
>P97c1	ACTTTTTTGTTGCAAAAAAAAAAAAAAAAAAGTATAATTTTAATTGTACAA	
>glyA	TTAAAAAAATTATTTTTTTGTTTTTTTAGTGTATAATGTGTAAAATTCCA	
>0225	AATAAAAAATAAAAAAAATTATTTTTTATGTTAAAATTATAATCGTAGGG	
>P97like	GGATTTTAGTTACTAAAAAATTAAAATATGGTATAATTTTAATTAAATTG	
>0279	GATTTTTTTTTAAAAATTTTTAAAAATAGTGTAAAATTGTTAAATTATGA	
>glpK	AATTTCACAGGGCTCCTTTGGATTAAAATGTTATAATTCAAATATATAAA	
>P37like	TAAAAAAAATTTATAATTCTTCCTTCTATGATATAATAATTTCAGGTAAA	
>efp	TTCATTTTTAGATTTTTTTTTTTTTTTATGCTATAATTTATAGTTACTTT	
>pgk	TGTTTTTTCTTAGTTTTTCAACTTAATAGTTTATAATATAACAAAAATAA	
>46K	TCATTTTTTAAAAAAAAATTGATTTTTATAGTATAATTTATTTGTATAAT	
>gyrA	GAAAATTCTTATTAAAACATAAAAAATATGGTATAATTTATACTTACCAC	
>pyrH	TTTTTTAATACATTTTTTTTCAAAAATAAAGTATAATAAAAGAAACTTTT	
>ktrA	GATAATTTTAAAAATTTTTTCAATTTTGGTCTACAATTTAGTCAAATGAA	
>rplJ	AAAAAATACTTTTTTTATTTTCGCTTTCTGGTATAATTCAAAAACGCAAT	
>dam	TTTTTTAAATAATTTTATCCCTATTTGCTTATATAATTTAGTTTATGTCA	
>leuS	ACTTTTGGCTTTTAATTTTAAAAAATTATGCTATAATTTAGGTAACTATC	
>P146	GAGAAATTTTTTTAATTTTTAACTTCTATAGTATAATTATTGTATCACTT	

9. PUBLICAÇÕES

9.1 Publicação resultante desta tese

- **Artigo publicado na revista DNA Research**

  (Fator de impacto JCR 2010: 4,75)

WEBER, S.S.; SANT'ANNA F.H. & SCHRANK, I.S. Unveiling *Mycoplasma hyopneumoniae* promoters: sequence definition and genomic distribution. DNA Research 19 (2): 103-115, 2012.

DNA Research pp. 1–13, (2012) doi:10.1093/dnares/dsr045

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 × *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*	ACTTCAAGATTAATTATACCAATTT	TTTG	T	TAAAAT	TATAATA	77	ATG
MHP7448_0101	*clpB*	ACTCTTTACTTTAAGTGCCAAAAAA	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*	TTAAAAAAAATATTTTTTTGTTTTTT	TAGT	G	TATAAT	GTGTAAAA	5	ATG
MHP7448_0225		AATAAAAAAAAAAAAAAATTATTTTT	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	TAAAAAAAATTATAATTCTTCCTTC	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*	AAAAAATACTTTTTTATTTTCGCTT	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.

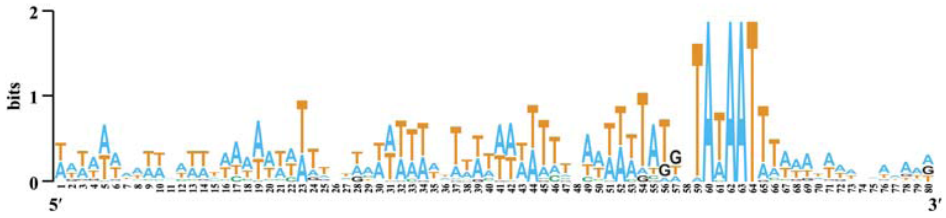


Figure 1. Sequence conservation in the *M. hyopneumoniae* promoter region. Sequence logo derived from the alignment of the 23 defined promoter regions showing the high conservation of the −10 element (positions 59–64), the presence of a semi-conserved −16 element (positions 54–57), the absence of a −35 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

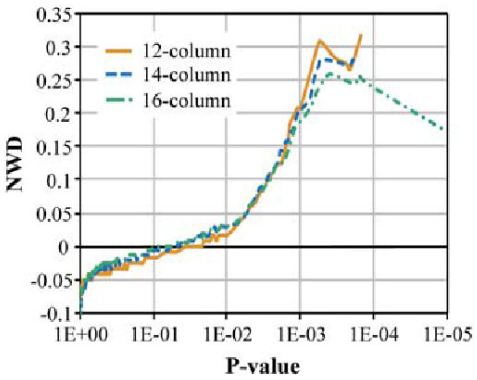


Figure 3. Performances of the 12-, 14- and 16-column PSSMs using a Markov order of 1 as the background model. Each curve shows the normalized weight score difference (NWD) calculated from theoretical and empirical distributions obtained for each matrix using a Markov order of 1. The higher the NWD value,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

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.

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

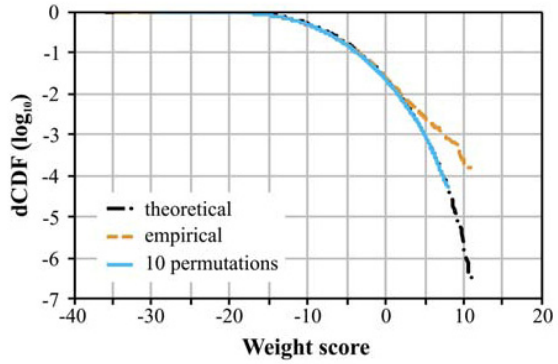


Figure 4. Weight score distributions for the 12-column PSSM. The curves of the theoretical (black; dashed-dotted line) and empirical (orange; dashed line) distributions obtained with the 12-column PSSM began to separate at a weight score of 3, indicating that the promoters were being distinguished from 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).

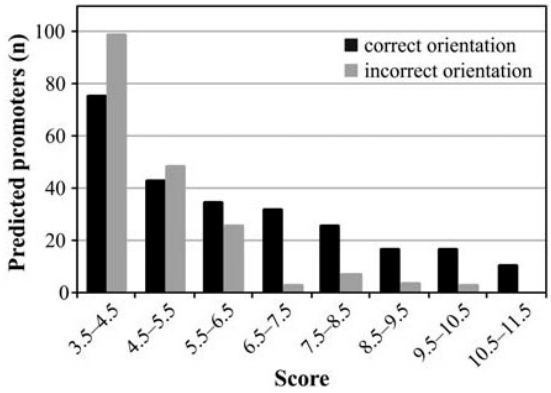


Figure 5. Weight score threshold definition. Distribution of the scores of the correctly and incorrectly oriented promoters predicted in the *M. hyopneumoniae* intergenic regions. Note that from score 6.5, the frequencies of incorrectly oriented promoters are much smaller than the frequencies of correctly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

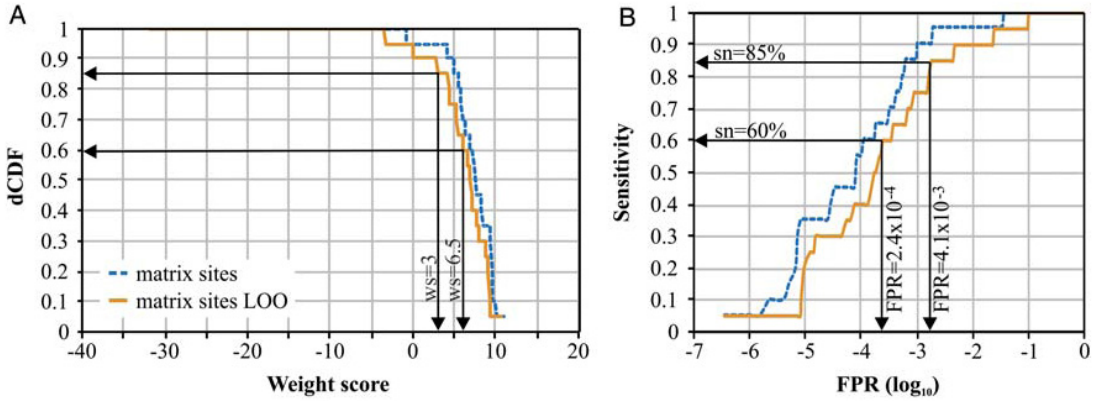


Figure 6. Trade-off between the sensitivity and FPR of the 12-column PSSM. (**A**) Score distributions of the experimentally defined sites used to build the matrix. Blue (dashed line), biased scores assigned by the matrix to the defined sites. Orange (solid line), unbiased scores obtained using the LOO 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

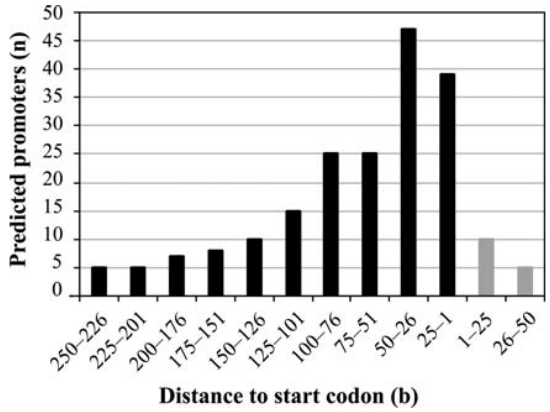


Figure 7. Distance of the predicted promoters from the annotated start codon of the *M. hyopneumoniae* CDSs. Distances were determined for the 201 predicted promoters scoring >6.5; they were measured from the −10 element to the start codon of the genes. Black bars indicate bases upstream of the start 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 (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.

9.2 Outras publicações

- **Artigo publicado na revista BMC Microbiology**

  (Fator de impacto JCR 2010: 2,96)

SANT'ANNA, F.H.; ANDRADE, D.S.; TRENTINI, D.B.; **WEBER, S.S**.; SCHRANK, I.S. Tools for genetic manipulation of the plant growth-promoting bacterium *Azospirillum amazonense*. BMC Microbiology (Online), v. 11, p. 107, 2011.

- **Artigo publicado na revista Journal of Molecular Evolution**

  (Fator de impacto JCR 2010: 2,31)

SANT'ANNA, F.H.; TRENTINI, D.B.; **WEBER, S.S.**; CECAGNO, R.; SILVA, S.C.; SCHRANK, I.S. The PII Superfamily Revised: A Novel Group and Evolutionary Insights. Journal of Molecular Evolution, v. 68, p. 322-336, 2009.

10. CURRICULUM VITAE

Dados pessoais

Nome Shana de Souto Weber

Formação acadêmica/Titulação

2007 Doutorado em andamento em Programa de Pós-Graduação em Biologia Celular e Molecular.
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.
Título: Determinação de regiões promotoras de *Mycoplasma hyopneumoniae*
Orientador: Irene Silveira Schrank.
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

2006 - 2007 Mestrado em Programa de Pós-Graduação em Biologia Celular e Molecular - UFRGS.
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.
Título: Fator σ de *Mycoplasma hyopneumoniae*: Mutagênese, Clonagem e Expressão
Ano de Obtenção: 2007.
Orientador: Irene Silveira Schrank.
Bolsista do(a): Coordenação de Aperfeiçoamento de Pessoal de Nível Superior.

2001 - 2005 Graduação em Ciências Biológicas.
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.
Título: Caracterização da região promotora do gene *vlhA* em diferentes isolados de *Mycoplasma synoviae*.
Orientador: Sérgio Ceroni da Silva.
Bolsista do(a): Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, Brasil.

Formação complementar

2006 - 2006 Clonagem e expressão em bactérias gram-positivas. (Carga horária: 16h).
Universidade Federal de Minas Gerais.

2004 - 2004 RDA em Microorganismos. (Carga horária: 16h).
Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.

Atuação profissional

1. Fundação Estadual de Produção e Pesquisa em Saúde.

Vínculo institucional

2005 - 2005 Vínculo: Estagiário, Enquadramento Funcional: Estágiario, Carga horária: 20

Atividades

2005 - 2005 Estágios, Lacen.
Estágio realizado
Desenvolvimento de método de detecção de *Chlamydia trachomatis* e Diagnóstico de *Mycobacterium* em amostras clínicas.

2005 - 2005 Atividades de Participação em Projeto, Centro de Desenvolvimento Científico e Tecnologico
Projetos de pesquisa
Detecção de *Chlamydia trachomatis* em amostras de urina masculina por reação em cadeia da polimerase

2. Universidade Federal do Rio Grande do Sul, UFRGS, Brasil.

Vínculo institucional

2007 - Atual Vínculo: Bolsista, Enquadramento Funcional: Doutoranda, Regime: Dedicação exclusiva.

Vínculo institucional

2006 - 2007 Vínculo: Bolsista, Enquadramento Funcional: Mestranda, Regime: Dedicação exclusiva.

Vínculo institucional

2002 - 2006 Vínculo: Bolsista iniciação científica, Enquadramento Funcional: Estagiária, Carga horária: 20

Atividades

03/2007 - Atual Atividades de Participação em Projeto, Centro de Biotecnologia, UFRGS.
Projetos de pesquisa
Determinação das regiões promotoras de *Mycoplasma hyopneumoniae*.

04/2006 - 12/2007 Atividades de Participação em Projeto, Centro de Biotecnologia, UFRGS.
Projetos de pesquisa
Clonagem e expressão do fator σ de *Mycoplasma hyopneumoniae* em *Escherichia coli*.

04/2005 - 12/2006 Atividades de Participação em Projeto, Centro de Biotecnologia, UFRGS.
Projetos de pesquisa
Resistência aos antimicrobianos e presença de plasmídeos em isolados clínicos e ambientais de *Escherichia coli*.

09/2002 - 04/2006 Estágios, Centro de Biotecnologia.
Estágio realizado
Iniciação Científica.

11/2004 - 12/2005 Atividades de Participação em Projeto, Centro de Biotecnologia, UFRGS.
Projetos de pesquisa
Análise da região controladora do gene vlha de *Mycoplasma synoviae*.

09/2002 - 09/2004 Atividades de Participação em Projeto, Centro de Biotecnologia, UFRGS.
Projetos de pesquisa
Clonagem e expressão do gene *apxIV* de *Actinobacillus pleuropneumoniae*.

Áreas de atuação

1. *Grande área:* Ciências Biológicas / *Área:* Genética / *Subárea:* Genética Molecular e de Microorganismos.

Prêmios e títulos

2006 Mensão honrosa, 25ª Reunião de Genética de Microrganismos.

Produção em C,T & A

Produção bibliográfica

Artigos completos publicados em periódicos

1. WEBER, S.S.; SANT'ANNA, F. H.; SCHRANK, I. S. Unveiling *Mycoplasma hyopneumoniae* Promoters: Sequence Definition and Genomic Distribution. DNA Research, v. 19, p. 103-115, 2012.

2. SANT'ANNA, F.H.; ANDRADE, D.S.; TRENTINI, D.B.; WEBER, S.S.; SCHRANK, I.S. Tools for genetic manipulation of the plant growth-promoting bacterium *Azospirillum amazonense*. BMC Microbiology (Online), v. 11, p. 107, 2011.

3. COSTA, M.M.; DRESCHER, G.; MABONI, F.; WEBER, S.S.; SCHRANK, A.; VAINSTEIN, M.H.; SCHRANK, I.S.; VARGAS, A.C. Virulence factors, antimicrobial resistance, and plasmid content of *Escherichia coli* isolated in swine commercial farms. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, v. 62, p. 30-36, 2010.
4. SANT'ANNA, F.H.; TRENTINI, D.B.; WEBER, S.S.; CECAGNO, R.; SILVA, S.C.I.; SCHRANK, I.S. The PII Superfamily Revised: A Novel Group and Evolutionary Insights. Journal of Molecular Evolution, v. 68, p. 322-336, 2009.
5. COSTA, M. M.; WEBER, S.S.; MABONI, F.; FERRONATTO, A. I; SCHRANK, I. S.; VARGAS, A. C. Patotipos de *Escherichia colina* suinocultura e seu impacto ambiental e na resistência aos antimicrobianos.. Arquivos do Instituto Biológico (Impresso), v. 76, p. 509-516, 2009.
6. COSTA, M. M.; DRESCHER, G.; MABONI, F.; WEBER, S.S.; BOTTON, S.A.; VAINSTEIN, M.H.; SCHRANK, I.S.; Vargas, A.C. Virulence factors and antimicrobial resistance of *Escherichia coli* isolated from urinary tract of swine in southern of Brazil. Brazilian Journal of Microbiology, v. 39, p. 741-743, 2008.
7. BROETTO, L.; CECAGNO, R.; SANT'ANNA, F.H.; Weber, S.; SCHRANK, I.S. Stable transformation of *Chromobacterium violaceum* with a broad-host-range plasmid. Applied Microbiology and Biotechnology, v. 71, p. 450-454, 2006.

Resumos expandidos publicados em anais de congressos

1. COSTA, M. M.; SILVA, M. S.; MABONI, F.; WEBER, S.S.; KOLLING, L.; DRESCHER, G.; VARGAS, A. C.; SCHRANK, I. S.. Resistência aos antimicrobianos e presença de plasmídeos em isolados clínicos e ambientais de Escherichia coli. In: XII Congresso Brasileiro de Veterinários Especialistas em Suínos, 2005, Fortaleza. Anais do XII congresso brasileiro da ABRAVES, 2005. v. 2. p. 7-8.

Resumos publicados em anais de congressos

1. WEBER, S.S.; SCHRANK, I.S. Análise do sítio de início de transcrição de genes de *Mycoplasma hyopneumoniae*: caracterização de promotores. In: 25º Congresso Brasileiro de Microbiologia, 2009, Porto de Galinhas - PE. 25º Congresso Brasileiro de Microbiologia, 2009.
2. GALVES, F. R.; WEBER, S.S.; SCHRANK, I. S. Construção de um vetor espécie-específico para transformação de *Mycoplasma hyopneumoniae*. In: XX Salão de Iniciação Científica e XVII Feira de Iniciação Científica da UFRGS, 2008, Porto Alegre - RS. Livro de Resumos - XX Salão de Iniciação Científica e XVII Feira de Iniciação Científica da UFRGS, 2008. p. 312.
3. REOLON, L.; WEBER, S.S.; LOPES, B. M. T.; SILVA, S. C.; SCHRANK, I. S. *In vivo* promoter activities in *Mycoplasma* hyopneumoniae: evaluation of a reporter system.. In: XI Congreso Argentino de Microbiologia, 2007, Córdoba. XI Congreso Argentino de Microbiologia, 2007.
4. REOLON, L.; WEBER, S.S.; SILVA, S. C.; SCHRANK, I. S. Atividade *in vivo* de promotores de Mycoplasma hyopneumoniae: avaliação de um sistema repórter. In: XIX Salão de Iniciação Científica e XVI Feira de Iniciação Científica da UFRGS, 2007, Porto Alegre - RS. Livro de Resumos - XIX Salão de Iniciação Científica e XVI Feira de Iniciação Científica da UFRGS, 2007. p. 456-457.
5. WEBER, S.S.; COSTA, M. M.; SILVA, M. S.; MABONI, F.; FERRONATTO, A. I.; DRESCHER, G.; VARGAS, A. C.; SCHRANK, I. S. Resistência aos antimicrobianos e presença de plasmídeos em isolados clínicos e ambientais de Escherichia coli. In: 25ª Reunião de Genética de Microrganismos, 2006, São Pedro - SP. Resumos da 25ª Reunião de Genética de Microrganismos, 2006. p. 102.
6. WEBER, S.S.; SCHRANK, I. S.; SILVA, S. C. Clonagem e expressão do fator σ de *Mycoplasma hyopneumoniae* em *Escherichia coli*. In: XVIII Salão e XIV Feira de Iniciação Científica, 2006, Porto Alegre - RS. Livro de Resumos XVIII Salão de Iniciação Científica e XV Feira de Iniciação Científica, 2006.
7. WEBER, S.S.; SCHRANK, I. S. Clonagem e expressão do fator σ de Mycoplasma hyopneumoniae em Escherichia coli. In: VIII Reunião Anual do Programa de Pós-graduação em Biologia Celular e Molecular do Centro de Biotecnologia da UFRGS, 2006, Porto Alegre - RS. Livro de Resumos PPGBCM, 2006. p. 133.
8. WEBER, S.S.; SCHRANK, I. S.; SILVA, S. C. Caracterização da região promotora do gene *vlhA* em diferentes isolados de *Mycoplasma synoviae*. In: 25ª Reunião de Genética de Microrganismos, 2006, São Pedro - SP. Resumos da 25ª Reunião de Genética de Microrganismos, 2006. p. 103.
9. REOLON, L.; WEBER, S.S.; SILVA, S. C.; SCHRANK, I. S. Caracterização funcional de promotores de *Mycoplasma hyopneumoniae*. In: XVIII Salão e XV Feira de Iniciação Científica, 2006, Porto Alegre - RS. Livro de Resumos XVIII Salão de Iniciação Científica e XV Feira de Iniciação Científica, 2006.

10. WEBER, S.S.; SCHRANK, I. S.; SILVA, S. C. Análise da região controladora do gene *vlhA* de *Mycoplasma synoviae*. In: XVII Salão e XIV Feira de Iniciação Científica, 2005, Porto Alegre, 2005.
11. WEBER, S.S.; SCHRANK, I. S.; SILVA, S. C. Análise da região controladora do gene *vlhA* de *Mycoplasma synoviae*. In: 51º Congresso Brasileiro de Genética, 2005, Águas de Lindóia, 2005.
12. BECKER, D.; JARDIM, A. F. M.; WEBER, S.S.; SANT'ANNA, F. H.; SANTOS, S.; RIBEIRO, M. O.; SCHERER, L. Contribuição da PCR in house no diagnóstico da tuberculose: a experiência em rotina de um laboratório de saúde pública. In: XXIII Congresso Brasileiro de Microbiologia, 2005, Santos - SP. XXIII Congresso Brasileiro de Microbiologia, 2005.
13. WEBER, S.S.; COSTA, M.M.; SILVA, S.C.; SCHRANK, I.S. O gene *apxIV* de *Actinobacillus pleuropneumoniae*: expressão e purificação da proteína em *E. coli*. In: XVI Salão e XIII Feira de Iniciação Científica, 2004, Porto Alegre. Livro de Resumos XVI Salão de Iniciação Científica e XIII Feira de Iniciação Científica, 2004.
14. WEBER, S.S.; COSTA, M.M.; SILVA, S.C.; SCHRANK, I.S. Clonagem e expressão em *E. coli* do gene *apxIVA* de *Actinobacillus pleuropneumoniae*. In: XXIV Reunião de Genética de Microorganismos, 2004, Gramado, 2004.
15. WEBER, S.S.; COSTA, M.M.; SILVA, S.C.; SCHRANK, I.S. Clonagem e expressão do gene *apxIV* de *Actinobacillus pleuropneumoniae* em *E. coli*.. In: XV Salão e XII Feira de Iniciação Científica, 2003, Porto Alegre, 2003.

Eventos

Participação em eventos

1. 25º Congresso Brasileiro de Microbiologia. Análise do sítio de início de transcrição de genes de *Mycoplasma hyopneumoniae*: caracterização de promotores. 2009. (Congresso).
2. 26ª Reunião de Genética de Microrganismos. Transcrição em *Mycoplasma hyopneumoniae*: análise do fator sigma da RNA polimerase e estratégias para a identificação de outros possíveis candidatos. 2008. (Congresso).
3. X Reunião Anual do Programa de Pós-graduação em Biologia Celular e Molecular do Centro de Biotecnologia da UFRGS. Análise do sítio de início de transcrição de genes de *Mycoplasma hyopneumoniae*. 2008. (Outra).
4. IX Reunião Anual do Programa de Pós-graduação em Biologia Celular e Molecular do Centro de Biotecnologia da UFRGS. Mutagênese, clonagem e expressão heteróloga do fator σ de *Mycoplasma hyopneumoniae*. 2007. (Outra).
5. 25ª Reunião de Genética de Microrganismos. Resistência aos antimicrobianos e presença de plasmídeos em isolados clínicos e ambientais de *Escherichia coli*. 2006. (Congresso).
6. XVIII Salão e XV Feira de Iniciação Científica. Clonagem e expressão do fator σ de *Mycoplasma hyopneumoniae* em *Escherichia coli*. 2006. (Outra).
7. VIII Reunião Anual do Programa de Pós-graduação em Biologia Celular e Molecular do Centro de Biotecnologia da UFRGS. Clonagem e expressão do fator σ de *Mycoplasma hyopneumoniae* em *Escherichia coli*. 2006. (Outra).
8. 51º Congresso Brasileiro de Genética. Análise da região controladora do gene *vlhA* de *Mycoplasma synoviae*. 2005. (Congresso).
9. XVII Salão e XIV Feira de Iniciação Científica. Análise da região controladora do gene *vlhA* de *Mycoplasma synoviae*. 2005. (Outra).
10. XIV Reunião de Genética de Microrganismos. Clonagem e expressão em *E. coli* do gene *apxIVA* de *Actinobacillus pleuropneumoniae*. 2004. (Congresso).
11. XVI Salão e XIII Feira de Iniciação Científica. O gene *apxIV* de *Actinobacillus pleuropneumoniae*: expressão e purificação da proteína em *E. coli*. 2004. (Outra).
12. XV Salão e XII Feira de Iniciação Científica. Clonagem e expressão do gene *apxIVA* de *Actinobacillus pleuropneumoniae* em *E. coli*. 2003. (Outra).